Current challenges in statistical DNA evidence evaluation

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof. mr. C. J. J. M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op 12 januari 2017
klokke 15 uur

door

Giulia Cereda
geboren te Milano, Italië
in 1988

Promotores:

Prof. dr. R. D. Gill (Universiteit Leiden)

Prof. dr. F. Taroni (Université de Lausanne)

Samenstelling van de promotiecommissie:

Prof. dr. A. W. van der Vaart (Universiteit Leiden, chairman)

Prof. dr. P. Grünwald (Universiteit Leiden, secretary)

Prof. dr. A. Biedermann (Université de Lausanne)

Prof. dr. J. Mortera (Università di Roma tre)

Prof. dr. M. Sjerps (Universiteit van Amsterdam)

This thesis is part of a "cotutelle" agreement between the universities of Lausanne and Leiden. It has been supported by the Swiss National Science Foundation, through grants no. 105311-1445570 and 10531A-156146.

Alla mia famiglia

The front cover of this book was drawn by my father, Paolo Cereda.
In the back cover Lord Ganesha, the god of wisdom and learning, the patron of arts and sciences, as well as the remover of obstacles. One of his two tusks was broken to write down a very important textbook, the Mahabarata. According to some myths, he was generated by his mother, Parvathi, alone, using turmeric paste. For this reason, he represents women's independence.

A catalogue record is available from the Leiden University Library.

Preface

DNA profiling has become one of the most widely used techniques for human identification in forensic science since its introduction in 1984 by Alec Jeffreys. Despite the common belief that DNA evidence is a "damning evidence" which leaves no space for uncertainty, it actually needs strong statistical models in order to be used as a support for particular conjectures. The process which allows forensic experts to evaluate the statistical meaning of DNA evidence is one of the most interesting domains of forensic science of the last decades. This thesis started with the aim of building a statistical interpretative framework for a new genotyping methodology, the DIP-STR marker system, conceived to deal with the problem of extremely unbalanced mixtures.

While working on this project, we were confronted with the so-called 'rare type match problem', a very interesting open problem of forensic DNA statistics. The term refers to the situation in which there is a correspondence between the DNA profile of a suspect and that of a recovered stain, but this profile was never observed in a previously collected reference database. The evaluation of such a correspondence is very challenging. This problem is very common when using Y-STR markers or new genotyping techniques, such as DIP-STR markers, since the coverage of the available databases is limited. Therefore, we started investigating several statistical methods to deal with the rare type match problem. This led to the in-depth study of other delicate methodological issues, such as uncertainty assessment, data reduction, hybrid solutions.

As a closing loop to this Phd project, one of the discussed methods is proposed as a solution to the DIP-STR rare type match problem.

Contents

Introduction

One of the main aims of forensic statistics is to evaluate to what degree some piece of evidence supports one or the other of the hypotheses of interest in a judicial setting. The largely accepted method to perform this evaluation is the calculation of the *likelihood ratio*, a statistic that expresses the relative plausibility of the observations under the hypotheses. For instance, a typical piece of evidence may be a trace, found at the crime scene, containing DNA material from a single donor and whose profile corresponds to a known suspect's DNA profile. A couple of mutually exclusive hypotheses is typically defined, of the kind of 'the crime stain came from the suspect' (h_p) and 'the crime stain came from an unknown donor' (h_d). The likelihood ratio is the ratio of the probabilities of observing the matching profiles under these two hypotheses.

In case of DNA mixtures (traces containing DNA from several contributors) common genotyping techniques do not allow to distinguish the DNA profile of each contributor. This complicates the statistical evaluation of this kind of evidence, since different combinations of DNA profiles are compatible with belonging to the (known and unknown) contributors to the stain. Moreover, using standard techniques, if the quantitative share of DNA of one of the contributors is less than 10% of the total DNA quantity, his DNA profile is generally 'masked' by the DNA profile of the other contributor(s). As a consequence, it is very difficult to detect this minor DNA with classical methods of genotyping. Unbalanced mixtures of this kind are quite common, for example in cases of sexual assaults when the victim's DNA is largely predominant. This means that there is a paramount need for reliable solutions.

The advent of a new technology, the DIP-STR (short for Deletion Insertion Polymorphisms - Short Tandem Repeats) marker system, constitutes an answer to the problem represented by the extremely unbalanced mixtures.

The initial aim of this thesis was to develop a Bayesian statistical model to evaluate DIP-STR results in the light of competing hypotheses of interest: this represents an essential element for rendering the potential of this new typing technique useful for practitioners. Furthermore, in this initial project we compared, from a statistical and forensic perspective, the usefulness and usability of the DIP-STR markers with that of traditional marker systems, such as classical STR and Y-STR markers.

While in progress, we were confronted with several delicate methodological issues regarding forensic statistics: first, we noticed that the Bayesian methods used in the literature can be seen as ad hoc approximation to the full Bayesian solution. Then, we were confronted with the so-called 'rare type match problem', the situation in which there is a match between

the characteristics of some recovered material and those of the control material, but these characteristics have not been observed yet in previously collected samples (i.e., they do not occur in any existing database of interest for the case). The statistical evaluation of such a scientific finding depends on the rarity of the characteristic of interest (such as a DNA profile) in the population of reference. Indeed, the rarer it is, the higher is the likelihood ratio. The uncertainty over this rarity is usually dealt with using the observed relative frequency of the profile in some available database, but in case of no occurrence existing solutions are not satisfactory. The rare type match problem is particularly significant when Y-STR (or mitochondrial) DNA profiles are used, and in presence of new genotyping techniques (such as DIP-STR markers), for which the available database size is still limited.

We decided to start working with Y-STR data to study both new and existing solutions for the evaluation of rare type matches: classical Bayesian methods (beta-binomial and Dirichlet-multinomial) were revisited and compared to a Bayesian nonparametric approach tailored explicitly for the rare type match problem. Two frequentist solutions are also analysed: the discrete Laplace method and a new solution based on the Good-Turing estimator.

While studying frequentist solutions, we realised that different methods are based on different reductions of data, and that this is seldom discussed in forensic literature. Moreover, frequentist solutions involve different levels of uncertainty which have to be considered and discussed. Working on both frequentist and Bayesian frameworks allowed us to better understand the difference between the two approaches, and between the full Bayesian approaches and the plug-in approximations found in the literature. A lemma is also developed to help obtaining the full Bayesian likelihood ratio.

As a closing loop for this project, one of the Bayesian methods developed as a solution for the rare type match problem is applied to DIP-STR data. Also, the evaluative model built for DIP-STR data, in the initial stages of the research, has been improved and extended to incorporate parameter uncertainty in a consistent Bayesian way.

Scope and Propositions

This thesis covers different problems concerning the evaluation of DNA evidence. It is mainly divided into two parts: the first regards the DIP-STR genotyping techniques. It addresses the imperative need of developing a model to assign the likelihood ratio for DIP-STR results, and compares, from a statistical and forensic perspective, the advantages of these novel set of markers compared to traditional marker systems, such as STR and Y-STR.

The second part deals with several more general statistical aspects involved in the evaluation of DNA evidence. It aims at defining the differences between full Bayesian methods and ad hoc plug-in approximations, and at solving the rare type match problem for Y-STR data. The issues of the different reductions of data and of the levels of uncertainty involved in frequentist solutions are also discussed.

These two parts are connected in the final project, by developing a Bayesian solution for the rare type match problem for DIP-STR marker system. Moreover, the initial model for

DIP-STR data is improved in the light of the statistical discussion of the second part: any ad hoc solution is avoided to obtain a full Bayesian approach.

Novelty

Extremely unbalanced mixtures are still a challenging area of DNA analysis. The DIP-STR marker system is a recently developed technique of genotyping, which has only been discussed from a biological point of view: when this PhD project started, no Bayesian statistical solutions for the likelihood ratio assessment of DIP-STR evidence was available, making this research innovative and extremely useful. Moreover, this research will take advantage of the use of graphical probability models, in particular OOBNs, because of their very intuitive and flexible structures. They provide a powerful language for constructing knowledge-based models for reasoning under uncertainty, and significantly simplify the calculation of the statistics of interest. This represents an improvement if compared to the formulaic approach, classically used in literature.

The rare type match problem is still an open challenge of forensic statistics: it is so important a problem that it has been called 'the fundamental problem of forensic mathematics' (Brenner, 2010). This research produced, compared, and discussed new methodologies (both Bayesian and frequentist) to address this problem. Some methods have been developed specifically with this purpose, others have been adapted from existing ones. The use of Bayesian nonparametric methods represents a novelty in forensic science. Lastly, a solution for the rare type match problem encountered with the DIP-STR marker system is provided and discussed.

An additional contribution of this thesis to the theoretical statistical framework of forensic science is the discussion about the distinction between Bayesian and frequentist approaches to likelihood ratio assessment. This is important inasmuch the ad hoc plug-in approximations can be seen as hybrid solutions between the two. Moreover, a new Lemma is introduced and proved. It is of very broad application, since it can be used in all those situations (very common in forensic science) in which one wants to obtain the Bayesian likelihood ratio for data that depends on parameters, when prosecution and defence agree on the distribution of part of the data, but disagree on the distribution of the rest of the data.

Outline

The core structure of this research is represented by a series of papers written during the five years of this doctoral research. The papers are:

- Cereda, G., Biedermann, A., Hall, D., and Taroni, F. (2014) "Object-oriented Bayesian networks for evaluating DIP-STR profiling results from unbalanced DNA mixtures" *Forensic Science International: Genetics,* Vol. 8, pag. 159-169.
- Cereda, G., Biedermann, A., Hall, D., and Taroni, F. (2014) "An investigation of the potential of DIP-STR markers for DNA mixture analyses" *Forensic Science International: Genetics,* Vol. 11, pag. 229-240.

- Cereda, G. (2016) "Impact of model choice on LR assessment in case of rare haplotype match (frequentist approach)" *Scandinavian Journal of Statistics,* In Press.
- Cereda, G. (2016) "Bayesian approach to LR for the rare match problem" *Statistica Neerlandica*, In Press.
- Cereda, G. "Nonparametric Bayesian approach to LR assessment in case of rare haplotype match" (submitted to *Annals of Applied Statistics*).
- Cereda, G., Gill, R. D., and Taroni, F. "A solution for the rare type match problem when using the DIP-STR marker system" (submitted to *Forensic Science International: Genetics*).

The dissertation is structured in three parts.

Part I
Chapter 1 contains the forensic and statistical knowledge essential to appreciate the results of this research, while Chapter 2 summarizes the results which constitute the content of each paper, and provides the logical thread that links the diverse studies.

Part II
Chapters 3 to 8 are each in the form of a separate research paper, in the order in which they are written.

Part III
Chapter 9 discusses the contribution and implications of the results in a larger statistical and forensic context, and describes further research directions and open questions.

Notation

Throughout Chapter 1 and 2, random variables and their values are denoted, respectively, with uppercase and lowercase characters: x is a specific realisation of X. Random vectors and their values are denoted, respectively, by uppercase and lowercase bold characters: $\mathbf{p}$ is a realisation of the random vector $\mathbf{P}$. Bayesian probability is denoted with $\Pr(\cdot)$, while density of a continuous random variable X is denoted by $p(x)$ or $f(x)$. For a discrete random variable Y, both the continuous notation $p(y)$ and the discrete one $\Pr(Y = y)$ will be used. Frequentist probability will be denoted as $\mathscr{Pr}$. Occasionally we deviate from this rule. For instance, when dealing with graphical models, the notation changes: `teletype` is used for classes, **bold** for nodes and *italic* for states. Chapters 3, 4, 5, 6, 7, 8 stand aside and have a distinct convention, mostly based on the journal requirement for each paper.

Part I

Background and summary

Chapter 1

Preliminary concepts

1.1 Forensic DNA analysis

The unicity of each person's entire DNA sequence makes of DNA traces one of the most useful type of scientific findings for forensic identification. This is why, from its first use in the UK in 1986 (*R vs Colin Pitchfork*, Wambaugh (1989)), its use in forensic applications has become widespread (Walsh et al., 2004).

The contents of the forthcoming subsections are mainly based on Buckleton et al. (2005), Butler (2005, chap. 5) and Coquoz and Taroni (2006). They describe, from a biological and technical point of view, what DNA is and how it is used in the forensic context.

1.1.1 DNA as identification tool

DNA is the molecule that encodes the genetic instructions for the development and functioning of all known living organisms. It has a double-strand structure, where each strand is made up of a sequence of four nucleobases: Adenine, Cytosine, Guanine and Thymine (usually denoted by four letters, A, C, G and T). Each base is attached to a sugar molecule and a phosphate molecule to form *nucleotides*, arranged in the two long strands to form a spiral, with the shape of a double helix. Bases of one strand pair up with bases of the other strand – A with T, and C with G – to form units called *base pairs*. The instructions encoded in DNA are stored as a code made up of a double sequence of about 3 billions base pairs where the order of the bases in the sequence determines the information available for building and maintaining an organism. Since the bases are always paired A-T and C-G, to refer to a particular portion of the DNA sequence it is sufficient to consider one strand, such as AATTGCCTTTTAAAAA.

A distinct portion of DNA which codes instructions for a particular body's need (mostly the creation of proteins) is called *gene*. The 32,000 genes present in the human DNA form the so-called *genome*. All nucleotides are not aligned on a single chain: they are organised in thread-like structures called *chromosomes*. Humans have 46 chromosomes which form 23 couples of *homologous chromosomes*, identical to one another in shape and size, one inherited

from the mother and one inherited from the father. One of these pairs is composed by the *sex chromosomes* which, among other functions, determine the sex of the individual. The other 22 pairs of chromosomes are called *autosomal chromosomes* and determine the rest of the body makeup and functions. The DNA code, or *genetic code*, is passed to the offspring through the paternal sperm and the maternal egg: the mother passes 23 chromosomes through her egg, the father passes 23 chromosomes through his sperm.

The entire sequence of the DNA is unique to each individual. The reason for this variability is due to *recombination*, the process by the two chromosomes of each parent exchange some portions of the DNA sequence each other, shown in Figure 1.1. The resulting chromosomes, built of parts from the two chromosomes of one parent, are different from each of the two original chromosomes of that parent. *Meiosis* (or gamete cell production) is the process during

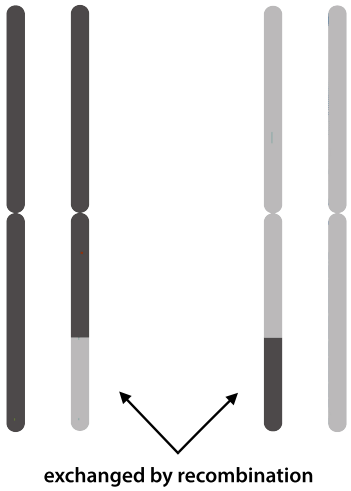


Figure 1.1: Genetic recombination between the chromosomes of each parent.

which each reproductive cell receives randomly one recombined chromosome. Other sources of variability among individuals are *mutations*, which are changes in the nucleotide sequence of the genome, due to deletions, insertions or metamorphosis of some nucleotides.

Although the entire *genome* is unique to each person, the greatest part of it is similar in all humans and only 0.1% characterizes the different individuals: the variations present in this minute portion of DNA are called *polymorphisms*. For forensic purposes, only those portions of the DNA sequence which are known to display a polymorphism are analyzed: these zones are called *genetic markers* (or *loci*) and the alternative possible variants which the DNA sequence displays in these zones are called *alleles*. An individual can have at most two alleles for the same polymorphism: the one carried by the paternal chromosome and the one carried by the maternal chromosome. If these two alleles are equal, the individual is said to be *homozygous* at the specific marker, otherwise he is said to be *heterozygous*. The couple of alleles present at an individual's genetic marker is called *genotype* and a combination of alleles at adjacent locations on a chromosome, inherited together, is called *haplotype*. A *DNA profile* is the combination of genotypes of multiple markers. While the entire DNA sequence is unique to each individual, there is the possibility that a DNA profile can be shared by two unrelated persons, with a (usually tiny) probability that decreases when the number

of loci are analysed. Moreover, a father and a son have the same DNA sequence in their Y-chromosome. Hence, DNA profiles obtained from this portion of DNA are shared by many people in the same population.

There are two different kinds of polymorphism, which are commonly analyzed:

- The *length polymorphisms* are located in particular portions of the DNA molecule where a particular sequence of nucleotides repeats itself many times. Each allele is represented by the number of such repetitions.
- The *sequences polymorphisms* are located in regions of the DNA strands where the type of one or more nucleotides varies among individuals. The different alleles are distinguished by a difference in one ore more nucleotides.

In this project *STR polymorphisms* (belonging to the first group), and *Deletion Insertion polymorphism* (belonging to the second group) will be used.

STR, Short Tandem Repeat Polymorphisms

A short tandem repeat (STR) polymorphism is a length polymorphism made of a pattern of two or more bases, which are repeated directly adjacent to each other. These patterns, called *words*, are typically repeated between 3 and 51 times: the polymorphism is represented by the difference in the number of repetitions of the same word, between the different individuals. *STR markers* are specific regions of the DNA in which an STR polymorphism is known to exist. Each STR allele is a number corresponding to the number of repetitions of the same sequence.

Below, the readers can see what the DNA sequence of an individual looks like, at the same locus of two homologous chromosomes, when an STR polymorphism is present.

Chromosome A_1
..CGGGT$\underbrace{\text{ATTGATTGATTGATTGATTGATTGATTGATTG}}_{8\times\text{ATTG}}$GAAAGGT..
Chromosome A_2
..CGGGT$\underbrace{\text{ATTGATTGATTGATTGATTGATTG}}_{6\times\text{ATTG}}$GAAAGGT..

The repeating pattern is the word ATTG. The genotype of this person at this locus is (6,8). Repeat numbers could be integers or decimals: for instance, if the repeat number is 9.3, this means that there are 9 repetitions and an incomplete repeat consisting of 3 more letters.

Insertion-deletion polymorphisms

Insertion-deletion (INDELs or DIPs) are length polymorphisms created by the presence or the absence of short (typically 1 to 50 base pairs) sequences of nucleotides in the human genome (Pereira et al., 2009a). DIPs are diallelic, the two possible alleles being L (for *long*) in case the specific combination of nucleotides is present, or S (for *short*) in case it is absent.

Allele L
..ATGCGT AATT TAGGGCTGGATC...

Allele S
..ATGCGTTAGGGCTGGATC.....

In the example, the sequence of nucleotides AATT is present in the first sequence, while it is absent in the second one.

In the last years, with the identification of 2000 human diallelic INDELs (Weber et al., 2002), this kind of polymorphisms has received major attention. Since then, a large literature has been developed, about genetic structure of human population (Yang et al., 2005; Rosenberg et al., 2005) and their use as genetic markers in natural population (Vali et al., 2008). A huge map of insertion-deletion variation in the human genome, which contains more than 415,000 distinct polymorphisms is published by Mills et al. (2006). DIPs are currently used for forensic purpose (da Costa Francez et al., 2012), especially for complex pedigree kinships (Pereira et al., 2009b) and for identification studies involving highly degraded DNA (Weber et al., 2002), a field in which the use of DIPs is very promising and more reliable than the use of STRs.

1.1.2 Technical steps of DNA genotyping

DNA can be found on different biological materials, such as blood, sperm, saliva, and hairs, but also on objects which have been in contact with human cells. After a trace is collected, its cells are broken down with chemical reagents, in order to reveal the DNA inside them. After this, the DNA is purified, with the aim of separating it from other molecules that can potentially interfere or inhibit the process of analysis.

In order to obtain a genetic profile from a DNA trace, it is necessary to amplify it across several orders of magnitude. This is done through a biochemical technology, called *Polymerase Chain Reaction* or *PCR* (Reynolds et al., 1991), which generates thousands of millions of copies of a selected portion of the DNA sequence. The method relies on about 30 cycles of repeated heating and cooling. The target DNA sequence to be amplified is pinpointed through the use of *primers*, which are short DNA fragments containing sequences of nucleotides complementary (according to the rule A-T, C-G) to the target region, together with an enzyme, the DNA polymerase, that adds the building blocks in the proper order based on the template DNA sequence.

The genetic markers to be amplified can be of different type. For this project STR markers, Y-STR markers and DIP-STR markers will be considered, as described in the forthcoming Sections 1.1.3 and 1.1.5.

A separation step is then required to pull the different targeted fragments apart, and to allow to distinguish the different alleles. This separation process is performed through *electrophoresis*, a technique which is used to separate molecules on the basis of their weight. The process is based on the migration undertaken by charged molecules, when immersed in a liquid and exposed to the electrical field generated by a couple of electrodes of opposite charge: negatively charged molecules move to the positive electrode, and vice versa. This movement has

a different speed, on the basis of the size of the molecules: light molecules will move faster than heavier ones. This causes the alleles to sort themselves according to weight. If applied to the amplified DNA fragments, the process consists of injecting the amplified sample into a gel, then to pass an electrical current through the gel, causing the alleles, which are all negatively charged, to move towards the positive pole. Since DNA is not visible in natural light, coloured dyes are used to monitor the progress of the electrophoresis. The output from the process is a graph, called *peak profile* or *electropherogram*, in which the horizontal axis gives the base pair measurement, and the vertical axis the light intensity. Each peak indicates the presence of an allele, where the height is a measure of the amount of the allele in the amplified sample. In Figure 1.2, an example of electropherogram obtained with STR markers is shown.

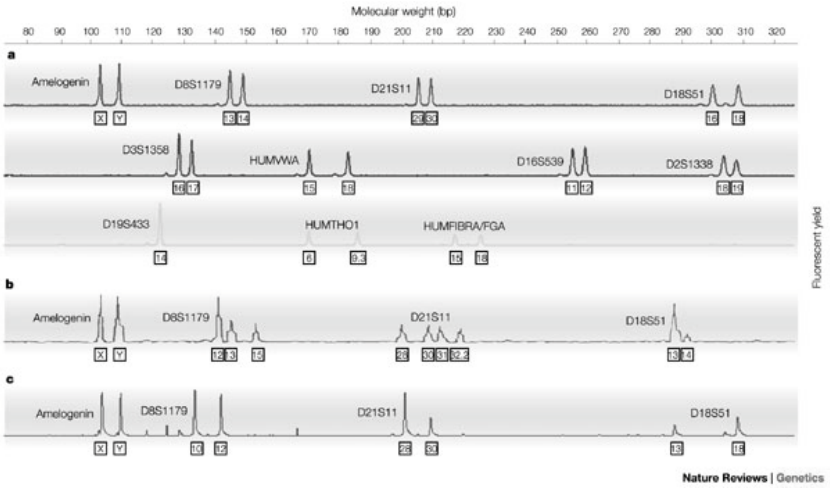


Figure 1.2: Electropherogram (Jobling and Gill, 2004) showing peaks at different *loci*. The numbers below each peak denote the corresponding allelic repeat numbers, while the name of the analysed *locus* is presented on the top of the peak(s) graph (i.e., D3S1358, vWA, etc.).

1.1.3 Two classical techniques of DNA genotyping

In forensic laboratories, two widely used marker systems are STR and Y-STR. This section provides an overview, highlighting advantages and drawbacks of the two methods.

STR marker system

An STR marker system is composed of a kit of analysis, designed to target some STR polymorphisms in a DNA trace of interest. These kits typically allow experts to target between 9 and 15 STR markers (plus Amelogenin, which is also used to assign the sex of the donor).

STR polymorphisms have been the predominant type of polymorphism used in human genetic studies since about 1990, and can be considered a standard approach to help addressing most

problems about forensic inference regarding identity of individuals (Chakraborty et al., 1999; Butler, 2011). There are several benefits in using STRs: they are multi-allelic and highly discriminating between unrelated and even closely related individuals, and they are relatively easy and not expensive to analyse (Vali et al., 2008).

Y-STR marker system

Y-STR markers are STR markers situated on the Y-chromosome (Butler, 2005). They are used in forensic casework (e.g. Roewer et al., 1992; Roewer, 2009), especially for their capacity to reveal male-specific Y-STR alleles in female/male DNA mixtures, even if extremely unbalanced (when classical STR markers are not performing adequately). These markers are thus very useful, in particular for cases of extremely unbalanced mixtures in which the major contributor is female and the minor one is male. There are however some limitations in the use of Y-STR markers. One the one hand, they are usable only for mixtures with a specific gender combination (Y-STRs only detect male component's DNA in a female background), reducing dramatically the number of suitable cases. On the other hand, a Y-STR profile can be quite common in a population (Vermeulen et al., 2009) and patrilineal relatives of a suspect cannot be excluded as being the contributors to the stain, if no mutations occur (Roewer, 2009). Recently, a panel of 13 rapidly mutating (RM) Y-STR markers has been identified (Ballantyne et al., 2012), which successfully differentiate between closely and distantly related males. However, they have the same gender restrictions of classical Y-STR markers.

It is also important to mention that, due to the lack of recombination, the different Y-STR markers form a single haplotype, formed by alleles that are not independent one another.

1.1.4 DNA mixtures

DNA mixtures are stains that contain genetic material from more than one person. This can be due to a contact between the different individuals' DNA material, anytime before the trace is collected.[1] Mixtures are commonly found, after sex rapes, in vaginal swabs obtained from the victim, as the traces usually contains material coming from the victim and the perpetrator (but also from other consensual partner(s)). A single contributor having at most two distinct alleles per locus, the main factor identifying a mixture is the presence of three or more alleles at at least one locus.

The criticality of mixtures is represented by the difficulty of the so-called "deconvolution", that is discerning the particular genotype of each contributor: this happens every time the different contributors share some alleles at some locus, as shown in the example of Table 1.1.

A way to separate the different contributors' allele, is to use information about the height (or the area) of the peaks. Some laboratories do so in order to distinguish the different contribu-

[1] Note also that a mixture can be generated after the collection of the trace, if contamination occurs.

Alleles detected in locus 1	**Alleles of the possible donors**		
11,12,13	(11,11) (12,13)	(11,12) (12,13)	(11,11) (11,12) (11,13)

Table 1.1: Three among the different combinations of the possible contributors' alleles, when (11, 12, 13) is observed at a specific locus. For the same detected alleles there may be a combination of homozygous-heterozygous contributors, two heterozygous contributors, or three contributors.

tors, based on their percentage of share in the whole mixture. This is called the *quantitative aspect* of the data, opposed to the *qualitative* one, which is retained for this research, and uses only information about the detection or not of the alleles at each locus.

1.1.5 Extremely unbalanced DNA mixture

One of the limitations of using STR markers is that this method does not work successfully if the proportion between the DNA quantities of the two contributors is more extreme than 1:10 (Clayton and Buckleton, 2005; Sutherland et al., 2009). Mixtures with these characteristics are referred to in this thesis as 'extremely unbalanced mixtures'. In the process of amplifying the DNA, the primers fail to pinpoint the alleles of the minor contributor, which are masked by those of the major one. Here, the threshold of 10% is retained as limit of detection of the minor DNA for blood:blood mixtures. This value varies depending on the type of biological fluids which constitute the mixture and on the specific combination of genotypes present in the mixture (Applied Biosystems, 2012), which should be assessed in the validation procedure (Butler, 2011).

Extremely unbalanced mixtures are quite common in forensic contexts, such as in cases of sexual assaults when the victim's DNA is largely predominant. Moreover, several fields of medical genetics are concerned, for example with the phenomenon of microchimerism during pregnancy, which is caused by the circulation of minute quantities (from 3% to 6%) of foetal DNA in the maternal blood (Lo et al., 1998; Tjoa et al., 2006), but also after organ transplant, when traces of the donor's DNA are present in the blood fluid of the transplanted patient (Gadi et al., 2006; Pujal and Gallardo, 2008). In forensic contexts, to address the constraint of these kind of mixtures, Y-STR markers are usually adopted, with the limitations described in Section 1.1.3.

As pointed out in Oldoni et al. (2015), it is difficult to estimate the precise incidence of unbalanced mixtures. Thus, it is undoubtely that most of the extremely unbalanced mixtures recovered so far have been evaluated under the assumption that they were single stains, losing interesting information about further contributors which they may have potentially provided. The solution to this problem is the use of the DIP-STR marker system, which allows the experts to obtain (at least part of) the minor contributor's genotype, in many cases, as explained in the next section.

DIP-STR marker system

As explained in Section 1.1.3, STR and Y-STR marker systems have some limitations: the first one is generally not working properly for extremely unbalanced mixtures, the second one requires a good gender combination between the two contributors to the stain and does not allow to discriminate between patrilinear relatives. Both the constraints of these methods can be overcome by the use of DIP-STR markers, which have recently been proposed as a novel type of genetic markers (Castella et al., 2013; Oldoni et al., 2015).

In fact, the problem of the masking of the minor genotype can be addressed with the use of primers that are allele-specific to assure that, each time the two contributors have different genotypes at some marker, the primers will anneal to different alleles. DIP markers have this characteristic: the primer specific for the allele S is different from the primer specific for the allele L.

However, the use of DIP markers alone, has the limitation of a low discriminating power. The novelty, here, consists in pairing a DIP polymorphism with a standard STR polymorphism, to increase the discriminating power and form a superlocus where the two component loci are not independent (less than 500bp apart).[2]

DIP-STR genotyping allows the selected amplification of the minor contributor's genotype as long as it has a DIP allele which is not in the major contributor's DIP alleles, in at least one marker. At each marker, the best scenario is when the DNA of the major and of the minor contributors are homozygous for different DIP alleles (i.e., one S-S and the other L-L). In this case, the possible results can show either two DIP-STR alleles of the minor contributor or one, depending on the STR-homozygosis or heterozygosis of the minor contributor. On the other hand, when the major contributor is DIP-homozygous and the minor contributor is DIP-heterozygous, only one haplotype of the minor DNA can be recovered (i.e., the one with the DIP allele different to the DIP allele of the major contributor's DNA). Table 1.2 summarizes the possible outcomes.

A limitation of this method is that, when the predominant DNA is DIP-heterozygous or both contributors are DIP-homozygous of the same type, it is not possible to obtain any result from the mixture: this happens because both the DIP primers (S and L), if used, anneal to the major contributor's DNA.

However, it is important to notice that, if the major contributor is DIP homozygous, some information about the minor contributor can be inferred by the absence of results (first and fifth row of Table 1.2): in fact, this absence indicates that the minor has the same DIP-homozygosity of the major (they are both S-S or L-L).

A first panel of 9 DIP-STR markers was first presented in Castella et al. (2013), while more recently 9 additional DIP-STR alleles have recently been made available (Oldoni et al., 2015). While within the same DIP-STR marker the DIP locus and the STR locus are chosen close enough so as not to recombine, independence can be assumed between the different allelic configurations of DIP-STR markers in the proposed panel.

[2]The two component loci are not independent because they are so close on the chromosomes that they cannot recombine.

DIP genotype of major contributor	DIP genotype of minor contributor	DIP-STR results using the DIP primer opposite to that of the major contributor
S-S	S-S	No results
	L-L	Complete genotype of the minor contributor
	S-L	Only the L DIP-STR allele
L-L	S-S	Complete genotype of the minor contributor
	L-L	No results
	S-L	Only the S DIP-STR allele
S-L	S-S L-L S-L	No results in each situation

Table 1.2: Informativeness of the different genotypic DIP-STR configurations.

Since DIP-STR alleles of the minor were successfully detected at ratios up to 1:1000 (Castella et al., 2013), the method reveals advantages for extremely unbalanced mixtures, for example in cases of sexual assaults when the victim's DNA is largely predominant, or cases of microchimerism during pregnancy. In Castella et al. (2013), the authors propose to test the use of DIP-STR markers also in early stages of pregnancy to perform kinship analyses, largely demanded for cases of pregnancy after rape (Guo et al., 2012).

1.2 Evaluation of DNA evidence

One of the main aims of forensic statistics is to evaluate to what degree some piece of evidence supports one or the other of exclusive hypotheses of interest. When the piece of evidence is a DNA trace which is found at the crime scene and whose profile corresponds to a known suspect's DNA profile, the hypotheses of interest are (unfortunately (Taroni et al., 2013)) often of the source level kind: 'the crime stain came from the suspect' (h_p) and 'the crime stain came from an unknown donor, unrelated to the suspect' (h_d).[3] When dealing with results from DNA mixtures of two contributors, the situation becomes more complicated, depending on the number of alleles observed at each marker. If the genotype of one of the two contributors (for instance, the victim of a sexual aggression) is known, and the suspect is compatible as contributor to the stain, the two hypotheses of interest become: 'the DNA in the mixed stain belongs to the victim and the suspect' (h_p), versus 'the DNA in the mixed stain belongs to the victim and to an unknown person, unrelated to the suspect' (h_d). The (unknown) true hypothesis h is the parameter of interest of our model. Usually the model involves also nuisance parameters, denoted as θ, unknown quantities necessary to perform the evaluation. To deal with the uncertainty over the nuisance parameters, additional data,

[3]This type of hypotheses are said 'source-level hypotheses' For a discussion on the use (misuse) of source level hypotheses, please refer to Champod et al. (2016). Note that this aspect is not considered in the current research.

which we will refer to as 'background', is usually given to the forensic statistician. This is partially different from the 'background information' I as defined in Aitken and Taroni (2004) and Taroni et al. (2014), but in many cases background data can be thought of as part of the background information. For instance, when dealing with DNA evidence, the nuisance parameter is often the list of the allelic frequencies in the population of interest, and the background data used to deal with it, is a sample of DNA profiles from the population in the form of a database. The reader is invited to notice the difference between θ and h: one is the parameter which we 'test' through the likelihood ratio (h), the other (θ) is a nuisance parameter involved in its calculation. Data to evaluate is made of evidence and background. The extent to which data is helpful to discriminate between the competing hypotheses of interest is called *probative value*.

1.2.1 Bayesian inference and likelihood ratio

Bayesian inference allows one to update subjective beliefs on propositions, when new information is gathered through Bayes' theorem. This is important for forensic statisticians that want to quantify how the available data modifies prior beliefs on hypotheses of interest (Lindley, 1977b; Taroni et al., 2010). Bayesian methods are divided into parametric (when parameters of the model are finite dimensional) or nonparametric (in presence of infinite dimensional parameters). The common ground is that a prior distribution is given to all the unknown quantities of the model. This prior should be based on the subjective belief of the Bayesian statistician performing the inference. This does not mean that these priors are arbitrary, rather they are based on the experience and set of knowledge of the individual, at a given time (Lindley, 1978).

The largely accepted method to quantify the probative value of given forensic findings (data, observation, measurement, ...) is the calculation of the *likelihood ratio*, a statistic that expresses the relative plausibility of the observations under the (generally) two hypotheses of interest (Robertson and Vignaux, 1995; Evett and Weir, 1998; Aitken and Taroni, 2004; Balding, 2005; Steele and Balding, 2014).

After a couple of hypotheses is given, the Bayesian likelihood ratio is defined as

$$\mathrm{LR} = \frac{\Pr(D = d \mid H = h_p)}{\Pr(D = d \mid H = h_d)}, \tag{1.1}$$

where Pr is the Bayesian probability, reflecting the expert's belief on the joint distribution of the random variables of the model, namely D (representing the data), H (representing the hypotheses), and Θ (the nuisance parameter(s)).

The likelihood ratio is used to quantify the way in which new information can change the belief, or 'odds', that a particular hypothesis is true. *Prior odds* and *posterior odds* are the odds before and after introducing information, such as a new piece of evidence. As part of Bayes' theorem, the likelihood ratio connects prior odds to posterior odds in the following way:

$$\underbrace{\frac{\Pr(H = h_p \mid D = d)}{\Pr(H = h_d \mid D = d)}}_{\textit{Posterior odds}} = \underbrace{\frac{\Pr(D = d \mid H = h_p)}{\Pr(D = d \mid H = h_d)}}_{\textit{Likelihood ratio}} \times \underbrace{\frac{\Pr(H = h_p)}{\Pr(H = h_d)}}_{\textit{Prior odds}}. \tag{1.2}$$

This formula clarifies that the likelihood ratio, which is a measure of the probative value of the findings D with respect to two alternative hypotheses, is to be distinguished from the conditional degree of belief on the same hypotheses (represented by posterior odds). The likelihood ratio is the ratio between posterior odds and prior odds. In a full Bayesian approach to the interpretation of evidence, the prior beliefs should be assigned by the commissioner (court, police) and should then be updated using the likelihood ratio, which is domain of the forensic laboratory, with the ultimate aim of obtaining the posterior beliefs. For a review on the importance of the likelihood ratio in the legal context, see Lindley (1991) and Aitken and Taroni (2004).

1.2.2 Frequentist likelihood ratio

On the other hand, in a frequentist context, the nuisance parameter θ and the hypotheses h are taken as fixed (unknown) quantities. The frequentist probability (here denoted as $\mathcal{Pr}$) can be expressed in terms of the Bayesian Pr, in the following way: $\mathcal{Pr}_\theta(\cdot \mid h) := \Pr(\cdot \mid \Theta = \theta, H = h)$, $\forall h$. The name is due to the fact that the probability of an event can be interpreted as the limit of its relative frequency in a large number of experiments. The frequentist likelihood ratio can be thus expressed as

$$\mathcal{LR}_\theta = \frac{\mathcal{Pr}_\theta(D = d \mid h_p)}{\mathcal{Pr}_\theta(D = d \mid h_d)}. \tag{1.3}$$

The difference between Bayesian and frequentist methods consists in how they treat the parameters θ and h. A Bayesian models the uncertainty about their value by random variables Θ and H, which are given prior distributions $p(\theta)$ and $p(h)$. Frequentists consider them as fixed (i.e., without distribution) unknown quantities.

One of the aims of this thesis is to carefully differentiate between the frequentist and the Bayesian approach, in order to provide guidelines for consistent solutions on both sides. A precise and concise account on problems and pitfalls in the LR definition and assessment can be found in Dawid (2016).

1.2.3 Rare type match problem

The evaluation of a match between the profile of a particular piece of evidence and a suspect's profile depends on the proportion of individuals with that profile in the population of potential perpetrators. Indeed, it is intuitive that the rarer the matching profile, the more the evidence is pointing against the suspect. Problems arise when the observed frequency of the profile in a sample from the population of interest (i.e., in a reference database) is 0. Such characteristic is likely to be rare, but it is challenging to quantify how rare it is. This problem is so

substantial that it has been defined "the fundamental problem of forensic mathematics" (Brenner, 2010).

The rare type match problem is particularly important in case a new kind of forensic evidence, such as results from DIP-STR markers is involved, and for which the available database size is still limited. The same happens when Y-chromosome (or mitochondrial) DNA profiles are used: because of the lack of recombination involved when offspring DNA is generated from the DNA of the parents, the haplotype must be treated as a unit and the set of possible haplotypes is extremely large. As a consequence, most of the Y-STR haplotypes are not represented in the database. This research will start by focussing on solving the rare type match problem for Y-STR markers. The final aim is to apply the studied solutions to DIP-STR markers.

1.2.4 Available solutions for the rare type match problem

The *empirical frequency estimator*, also called *naive estimator* or *maximum likelihood estimator* (MLE), that uses the frequency of the characteristic in the database, puts unit probability mass on the set of already observed characteristics, and it is thus unprepared for the observation of a new type. A solution could be the *add-constant* estimators (in particular the well-known *add-one* estimator, due to Laplace (1814), and the *add-half* estimator of Krichevsky and Trofimov (1981)), which add a constant to the count of each type, included the unseen ones. However, this method requires to know the number of possible unseen types, and it is also not very well-performing when this number is large compared to the sample size (see Gale and Church (1994) for an additional discussion).

Alternatively, Louis (1981) proposes the so-called 'rule of three', that states that if n is the size of the database, $3/n$ is a good approximation of the 95% upper bound for the frequency. This is also proposed in a Bayesian framework, by Jovanovic and Levy (1997); Winkler et al. (2002); Chen and McGee (2008).

Of interest for this research is the nonparametric *Good-Turing estimator* of Good (1953), based on an intuition on A. M. Turing. It is an estimator for the total unobserved probability mass, based on the proportion of singletons in the sample. If compared to the maximum likelihood estimator, or to the add one estimator, it has the advantage of being usable for the unobserved species, without additional constraints (Orlitsky et al., 2003). However, it does not allow to separate the frequencies of the unseen species, nor to estimate their number. For a comparison between *add one* and *Good-Turing* estimator, see Orlitsky et al. (2003).

As stated in Anevski et al. (2013), the *naive estimator* and the *Good Turing estimator* are in some sense complementary: the first gives a good estimate for the observed types, and the second for the probability mass of the unobserved ones. Lastly, the *high profile estimator*, introduced by Orlitsky et al. (2004), extends the tail of the *naive estimator* to the region of unobserved types. This estimator has been improved by Anevski et al. (2013) that also give the consistency proof.

Literature provides some examples of approaches to evaluate the likelihood ratio for the rare type match problem for Y-STR haplotypes: Egeland and Salas (2008), the κ method

Brenner (2010, 2014), the coalescent theory method (Andersen et al., 2013a), the haplotype surveying method (Roewer et al., 2000; Krawczak, 2001; Willuweit et al., 2011), and the discrete Laplace method (Andersen et al., 2013b). The latter was not proposed directly for the rare haplotype match case but is usable for that purpose. For more details about the discrete Laplace method, see Section 1.2.5.

Bayesian nonparametric estimators for the probability of observing a new type have been proposed by Tiwari and Tripathi (1989) using Dirichlet process, by Lijoi et al. (2007) using general Gibbs prior, and by Favaro et al. (2009) with specific interest to the two-parameter Poisson Dirichlet prior. However, for the LR assessment one has to obtain not only the probability of observing a new species but also the probability of observing this same species twice (according to the defence, the profile of the crime stain and the profile of the suspect are two independent observations).

1.2.5 The discrete Laplace method

The diversity of STR alleles in the population is the result of mutations. The *stepwise mutation model* (Kimura and Ohta, 1978) assumes that each mutation can, with equal probability, increase or decrease an STR repeat number by at most one, in one generation. In Caliebe et al. (2010), a Markov chain description of the stepwise mutation model is proposed, which shows nice convergence properties. The proposed process is called *normalized allele process* and models the difference between the STR alleles of the each individual and a fixed individual, in each generation. This process is proved to be a positive recurrent irreducible Markov chain on $\mathbb{Z}^{N-1}$. As such, it converges exponentially fast to the unique invariant distribution, which is unimodal if the mutation rate $\mu \leq 0.8$.

In Andersen et al. (2013b), the discrete Laplace distribution (Inusah and Kozubowski, 2006) is suggested (and empirically validated) as an approximation to the invariant distribution of the normalized allele process. A discrete random variable D is said to follow the discrete Laplace distribution $\mathrm{DL}(p)$, with $0 < p < 1$ if

$$\Pr(D = d) = f(d) = \Big(\frac{1-p}{1+p}\Big)p^{|d|}, \qquad \forall d \in \mathbb{Z}.$$

This result is used to model the distribution of single locus alleles, where the allele of reference which normalizes the process is estimated through the median of all the alleles at that locus.

According to this choice, the distribution of each single locus alleles X_i has density function

$$f(|d-m|) = \big(\frac{1-p}{1+p}\big)p^{|d-m|},$$

where m and p can be estimated using MLE from a sample $\{d_i\}_{i=1}^N$, as

$$\hat{m} = \mathrm{median}\{d_i\}_{i=1}^n,$$
$$\hat{p} = \hat{\mu}^{-1}\Big(\sqrt{\hat{\mu}^2+1} - 1\Big),$$

where

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} |d_i - \hat{m}|.$$

Let $\mathbf{X} = (X_1, X_2, ..., X_r)$ denote the random variable which describes an r-loci haplotype configuration. Moreover, there may be c different subpopulations to take into consideration. By making the strong assumption of independence between loci, within the same subpopulation, the following density is used to describe the probability that $\mathbf{X} = \mathbf{x}$:

$$f(\mathbf{x} \mid \{\mathbf{y_j}\}_j, \{\mathbf{p_j}\}_j) = \sum_{j=1}^{c} \tau_j \prod_{k=1}^{r} f(x_k \mid y_{jk}, p_{jk}).$$

For each j, τ_j is the probability a priori of generating from the jth subpopulation, while $\mathbf{p_j} = (p_{j1}, p_{j2}, ..., p_{jr})$ and $\mathbf{y_j} = (y_{j1}, y_{j2}, ..., y_{jr})$ represent the dispersion and location parameters, respectively, of the jth subpopulation.

The R package `disclapmix` (Andersen, 2013) allows one to estimate all the parameters of the model using the EM algorithm (Dempster et al., 1977), with initial subpopulation centres chosen via PAM algorithm (Kaufman and Rousseeuw, 2009), while their number is chosen by the BIC criteria (Schwarz, 1978).

1.3 Graphical models

The well-known *probabilistic graphical models* are graph-based representations that consist of a qualitative part, where features from graph theory are used, and a quantitative part that specifies the probability distributions over the nodes of the graph. They are defined as a "marriage between probability theory and graph theory" by Jordan (1998, 2004). Such network structures are formulated in a graphical communication language and can be seen as a compact representation of (conditional) dependencies and independencies between variables.

The most famous type of graphical models are the Bayesian networks, described in details in Section 1.3.1. The main goal of such models is to use conditional independences to find factorizations of the distribution over the graph structure, with the ultimate aim of computing efficiently the probabilities of interest.

Bayesian networks were first introduced by Pearl (1982), but other detailed accounts can be found in specialized literature (Pearl, 1988; Neapolitan, 1990; Jordan, 1998; Jensen and Nielsen, 2007; Cowell et al., 2007a). They are used in many fields where reasoning under uncertainty plays a central role (Pearl, 1988; Neapolitan, 1990; Gómez, 2004; Pourret et al., 2008) and they are thus becoming a more and more regularly used approach for analyzing problems in forensic science, where they are now part of well established literature (Aitken and Gammerman, 1989; Dawid and Evett, 1997; Dawid et al., 2002; Garbolino and Taroni, 2002; Taroni et al., 2004, 2014; Cowell et al., 2006b; Biedermann, 2007; Fenton and Neil, 2012). For all these reasons, Bayesian networks (and their object-oriented extension described in Section 1.3.2) are retained as the general modeling framework in this research.

The evaluation of DNA results via the likelihood ratio (described in Section 1.2), requires the calculation of the likelihood for data given the alternative hypothesis, which may be challenging. When adopted for kinship analyses, for example, likelihood ratio formulae become considerably complex, depending on parameters such as the supposed degree of relatedness and the number of individuals one needs to account for. Moreover, formulae may vary according to the genotypic configurations of the target individuals and the chosen genotyping technique. This computational burden can − as shown by foundational works of Dawid et al. (2002) − be approached and safely handled through Bayesian networks to obtain the same results as those obtained by Essen-Möller's formulaic approach (Essen-Möller, 1938). In fact, Bayesian networks allow one to obtain numerator and denominator of the likelihood ratio in few simple steps. Moreover, they prove to be a highly versatile framework that can accommodate analysts and reasoners with differing inferential interests (Taroni et al., 2014; Kjærulff and Madsen, 2008).

1.3.1 Bayesian networks: formal definition

A directed acyclic graph (DAG) is a graph formed by a collection of vertices, and by directed edges connecting one vertex to another, in a way that makes it impossible to start at some vertex v and follow a sequence of edges that eventually loops back to v again. Under its formal definition (e.g. Jensen and Nielsen, 2007), a Bayesian network is a DAG, whose vertices (also called nodes) form a finite set V and correspond to random variables $\mathcal{X} = \{X_v,\ v \in V\}$, while directed edges represent dependencies between variables. A node v is called a *parent* of a node w if there is a directed edge from v to w in the graph. Each node has a finite set of states and is equipped with a conditional distribution (given the parent nodes). For any ordering $X_1, X_2, ..., X_n$ of the random variables in $\mathcal{X}$, the edge set specifies the following factorization of the joint density p:

$$p(x_1, ..., x_n) = \prod_{i=1}^{n} \Pr(x_i | \mathrm{Pa}(x_i)), \tag{1.4}$$

where $\mathrm{Pa}(x_i)$ represents the set of the parent nodes of the node x_i. In case a node has no parent nodes, by $p(x_i|\emptyset)$ we mean $p(x_i)$. Equation (1.4) is called a *recursive factorization of p according to the DAG*, and formally defines Bayesian networks.

Bayesian networks are typically used for probabilistic inference. Each node represents a proposition or an assertion, such as those that an individual forms during the reasoning task. The probabilistic inference amounts to update the degree of belief on the truth of a particular proposition in the light of new information. Practically, the network is used to update knowledge about the state of a subset of variables when other variables are observed. The use of a factorization as that defined by Equation (1.4) reduces the computational effort of the procedure. In forensic applications, Bayesian networks are used to update the probabilities of the hypotheses of the prosecution (h_p) and of the defence (h_d) when relevant data for the case is obtained, but also the other way around: to update the probabilities of observing the data, under the two hypotheses h_p and h_d, (i.e., to calculate numerator and denominator of the

likelihood ratio). For this research we used the software Hugin,[4] specifically designed to work with Bayesian networks, along with the R package `RHugin`, which allows one to integrate the two platforms. A simple example of Bayesian network is the one for reasoning about diseases and symptoms, represented in Figure 1.3. Also known in specialized literature as the 'Classical diagnostic problem', it refers to a situation in which particular symptoms can arise as a consequence of different causes (or, diseases). This is a hypothetical, but widely applicable, scheme of representation, which is retained here for the sole purpose of remaining on a general level of discussion.

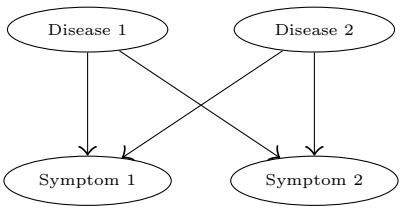


Figure 1.3: Bayesian network for generic reasoning about diseases and symptoms. Each node is Boolean with the state *True* representing the presence of the described disease or symptom, and the state *False* representing its absence.

The network models a situation in which there are two possible diseases and two symptoms. Both the diseases and the symptoms are not mutually exclusive. Each node is Boolean and the associated CPTs are shown in Tables 1.3 and 1.4.

Disease 1 = *False*	0.9
Disease 1 = *True*	0.1
Disease 2 = *False*	0.3
Disease 2 = *True*	0.7

Table 1.3: CPTs of the disease nodes. These two tables express the view that, initially, there is a probability of 0.1 for an individual to have one disease, and a probability of 0.7 to have the other.

Disease 1 =	*False*		*True*	
Disease 2 =	*False*	*True*	*False*	*True*
Symptom 1 = *False*	1	0.7	0.4	0.1
Symptom 1 = *True*	0	0.3	0.6	0.9
Disease 1 =	*False*		*True*	
Disease 2 =	*False*	*True*	*False*	*True*
Symptom 2 = *False*	1	0.9	0.4	0.05
Symptom 2 = *True*	0	0.1	0.6	0.95

Table 1.4: CPTs of the nodes representing the two symptoms. These tables contain the probability that the analyst assigns to the presence of the symptoms, given each possible configuration of the two disease nodes.

This network allows an analyst to revise his beliefs about the different diseases given the presence (or, absence) of one or both symptoms. For example, the analyst may ask questions

[4] http://www.hugin.com.

of the following kind: "If symptom 2 (but not symptom 1) is observed, what should be the probability that disease 1 affects the patient X?", "Does symptom 2 allow me to discriminate between the two 'causes', disease 1 and disease 2?", or "What if symptom 1 is absent?".

1.3.2 Object-orientation

The Bayesian network formalism offers interesting modeling capacity, but it is not always efficient or straightforward in its process of manual construction. For example, in case the model is composed of the repetitive use of some submodels, the 'copy and paste' system may be demanding, because all the submodels have to be updated to integrate new information (or evidence). In order to deal with this problem, object-oriented Bayesian networks (OOBN) can be employed (Koller and Pfeffer, 1997; Laskey and Mahoney, 1997; Bangsø and Wuillemin, 2000; Kjærulff and Madsen, 2008; Korb and Nicholson, 2011).

An object-oriented Bayesian network (also called a *class*) is a network that, in addition to regular nodes, contains *objects*. These are subnets that encapsulate themselves multiple objects (or subnetworks), giving rise to a composite hierarchical structure. Objects become part of a class under the form of so-called *instance nodes*.

One of the main advantages of using object-oriented Bayesian networks is that they are well suited for problem domains containing repetitive patterns. This happens typically when dealing with DNA evidence, in particular with DNA mixtures: in this case several contributors are represented in the network, each with a similar network substructure: with classical Bayesian networks, one is forced to use repetitive network fragments. With OOBNs, the use of instance nodes simplifies the problem, since the subnet is built only once, and then integrated in the class of interest through instances.

Each instance node is connected to other nodes of the 'external' class through the so-called *interface nodes*, divided into *input* and *output nodes*. As the name suggests, input and output nodes are the only nodes which directly connect any instance of the class to the external network: to the connected node(s), they hold the role of children or parent, respectively. Input nodes are actually placeholders for their parent nodes in the external network: they have the same states and the arrow only serves the technical purpose of establishing a logical equivalence. Throughout this text, instances are drawn as rounded rectangles, while interface nodes are grey: input nodes have a dashed outer border whereas output nodes have a solid line.

To show the utility of OOBNs one can consider again the generic problem of reasoning about diseases and symptoms, introduced earlier in Section 1.3.1. Imagine that the analyst wants to describe the progression in time of this setting. To do this with a Bayesian network, one possibility is to use a network structure as shown in Figure 1.4, where the same substructure is repeated twice. If several such periods need to be represented, this way of building the network can be time-demanding, especially if one needs to modify the CPT in the nodes of each repeated structure. With object-oriented Bayesian networks, one can create a class to model a single period, and use it in the form of instance nodes, where input nodes **Previous Disease 1** and **Previous Disease 2** are designed to be bound to nodes **Disease 1** and

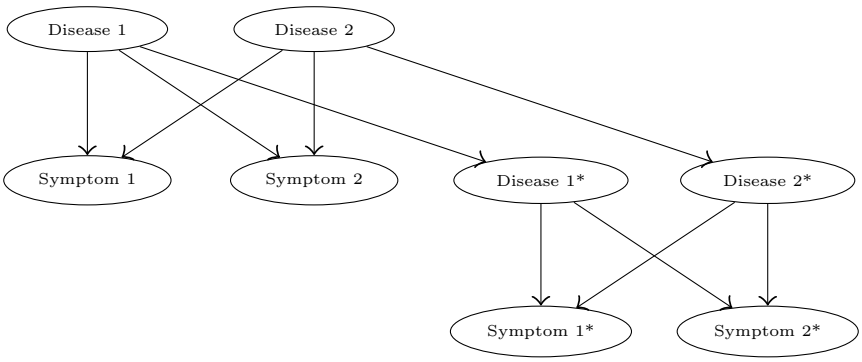


Figure 1.4: Bayesian network for reasoning about diseases and symptoms over time. Nodes **Disease 1** and **Disease 2** refer to the first period, while nodes **Disease 1*** and **Disease 2*** refer to the second period.

Disease 2 in the instance representing the previous period of time. This class is named `Period`, and is shown in Figure 1.5.

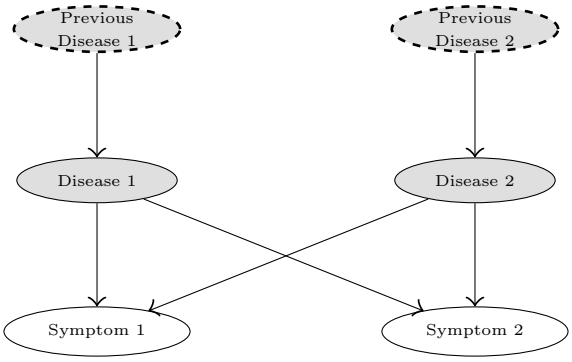


Figure 1.5: Representation of the class `Period`.

A model covering three (or more) periods can now be built simply by creating three instances of the class `Period`, as shown in Figure 1.6. As can be seen, nodes **Symptom 1** and **Symptom 2** do not appear in the external network, since they are not interface nodes.

Generally, an object-oriented Bayesian network can be used in the same way as a Bayesian network, instantiating some nodes to obtain the conditional probabilities of other nodes of interest.

Existing forensic literature considers the object-oriented Bayesian networks a particularly useful approach for addressing many evaluative aspects that are associated with evaluation of complex patterns of evidence (Dawid et al., 2007; Hepler et al., 2007; Taroni et al., 2014), or of results of DNA profiling analyses including mixtures, mutations, inference of source or kinship analyses. This is the reason why we chose OOBNs as the relevant methodological choice for the problem of interpreting DIP-STR results, which will be studied in this research.

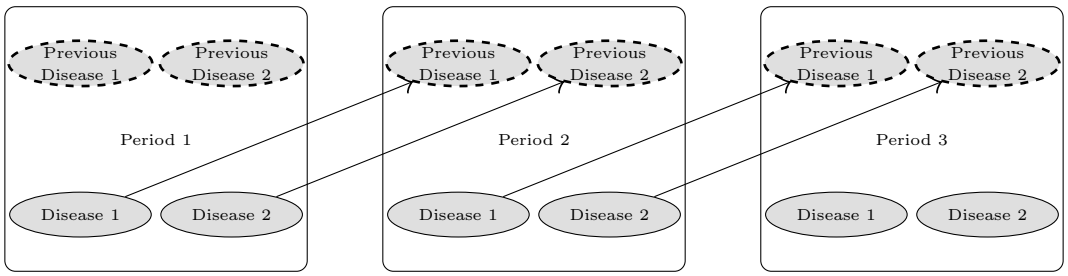


Figure 1.6: Expanded representation of an OOBN with three periods for reasoning about diseases and symptoms.

1.3.3 Bayesian networks in forensic DNA literature

The paper of Dawid and Evett (1997) is the first that introduces the use of Bayesian networks for DNA evidence evaluation, through an example involving blood stains.

Basic structures to deal with some aspects of DNA analyses have been proposed in the literature of the last decade.

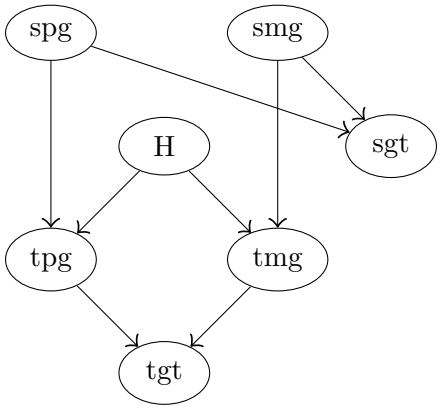


Figure 1.7: Bayesian network used for the evaluation of DNA results obtained from a crime stain (not mixed) and a suspect's stain.

Figure 1.7 shows a simple example of a Bayesian network to evaluate the correspondence between the DNA profile of a suspect and of a crime stain, focusing on individual genes and genotypes, presented in Dawid et al. (2002), and further discussed in Taroni et al. (2006). **H** is the hypothesis node, regarding whether or not the suspect is the source of the stain; **sgt** and **tgt** represent the genotype of the suspect and of the crime stain, respectively. Similarly, **spg** and **smg** represent paternal and maternal gene of the suspect, while **tpg** and **tmg** represent the paternal and maternal gene of the donor of the trace.

Another structure, which deals with the possibility of finding small quantities of DNA is proposed in Evett et al. (2002). A further important contribution in this context has been provided by Dawid et al. (2002), where the authors show the passage from initial pedigree representation of forensic identification problems to appropriate Bayesian networks representation.

A classical structure for the evaluation of DNA mixtures is shown in Figure 1.8 (Mortera

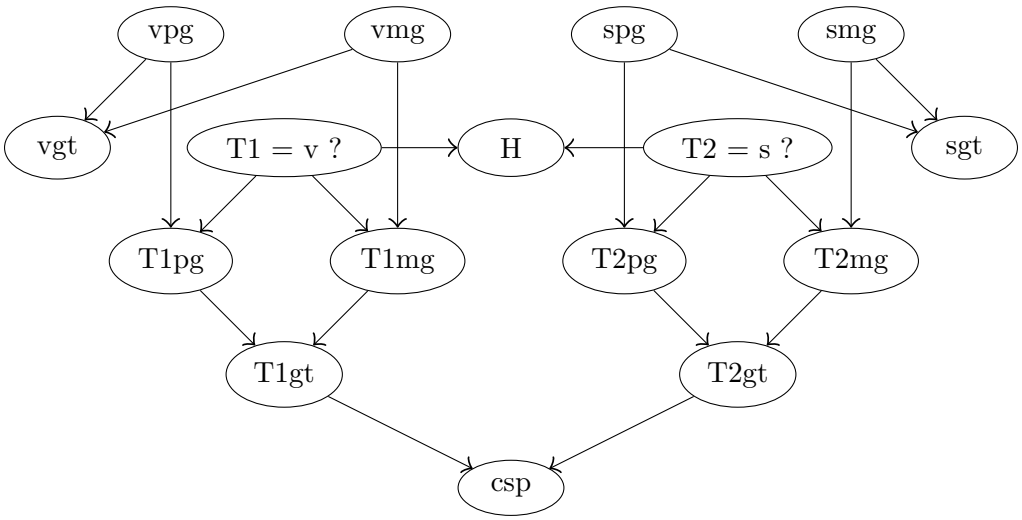


Figure 1.8: Bayesian network used for the evaluation of DNA results from a mixed DNA stain, recovered on a crime scene (Mortera et al., 2003).

et al., 2003). The network can be used to evaluate results from a crime stain (represented by **csp**), which contain two or more alleles per marker. **H** represents the hypotheses regarding the contributor of the stain, which may be (i) the suspect and the victim, (ii) the victim and an unknown individual, (iii) the suspect and an unknown individual, and (iv) two unknown individuals. The labels 'T1' and 'T2' denote, respectively, the first and the second contributors to the mixture. Nodes **vpg**, **vmg**, **spg** and **smg** represent the paternal and maternal genes of the victim and of the suspect, respectively, while **vgt** and **sgt** represent the victim's and the suspect's genotypes. Boolean nodes **T1=v?** and **T2=s?** model whether the victim has contributed to the crime stain or not, and whether the suspect has contributed to the crime stain or not, respectively. Nodes **T1pg**, **T2pg**, **T1mg** and **T2mg** represent their paternal and maternal genes, while **T1gt** and **T2gt** their genotypes.

Other more detailed structures are proposed in Mortera et al. (2003) and Biedermann et al. (2011b), the latter handling situations for which the number of contributors is not available. In Cowell et al. (2011), a network which models results from a quantitative approach and takes into account allelic drop-outs, stutters and silent alleles is presented. A specific overview of scientific literature about the use of Bayesian networks in forensic DNA applications can be found in Biedermann and Taroni (2012).

Bayesian networks have also been used for kinship analyses (Dawid et al., 1999, 2002), but more has been developed using object-oriented Bayesian networks (e.g. Hepler and Weir, 2008).

1.3.4 Object-oriented Bayesian networks for DNA evidence

It is in recent years that considerable research has been devoted to the application of object-oriented Bayesian networks to the evaluation of DNA evidence. As already mentioned, they are a very valuable tool to be used in this field, with different main perspectives, due to their versatility and modularity. Consider a model that describes the genotype, at a certain

marker, of two different persons: a suspect and an unknown person. The mechanism with which each of their alleles is inherited from their parents is the same for both individuals. Hence, in a Bayesian network it would give rise to four similar substructures (one for each allele). Through the use of an OOBN, it is possible to achieve this by using only one class, invoked in the network via four instances.

Another useful feature of OOBNs is their flexibility. In particular, a given network structure can readily be adapted to model other markers, or other individuals (such as the sibling of a missing suspect). As examples, an OOBN proposed in Green and Mortera (2009) is explained in details here (see Figure 1.9), to be used to evaluate DNA mixture evidence.

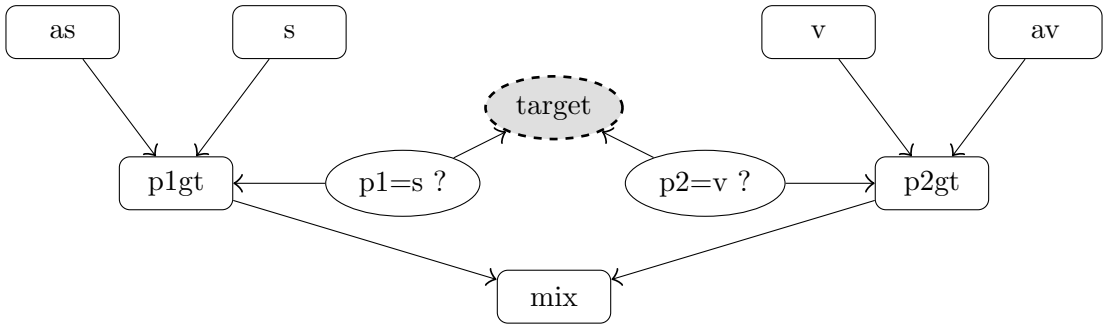


Figure 1.9: The network `Mixture` to be used for the evaluation of results obtained from a DNA mixture of two contributors.

The network `Mixture` of Figure 1.9 is used to calculate the strength of the evidence regarding a mixture of two contributors, the first of which can be either the suspect or an unknown person, while the second can be either the victim or an unknown person.

This network is made of instances of the classes `Genotype` and `Trace`, shown in Figure 1.10 and 1.11. The class `Genotype`, represented in the class `Mixture` by the instances **s**, **as**, **v**, and **av** is designed to represent the genotype, at a specific marker, of each individual.

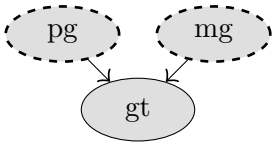


Figure 1.10: The class `Genotype`. The nodes **pg**, **mg** and **gt**, represent the paternal allele, the maternal allele, and the genotype itself, which is a logical combination of the parents' alleles.

The class `Trace` is represented in the class `Mixture` by the instances **p1gt** and **p2gt**. Ordinary nodes **p1=s?** and **p2=v?** of the class `Mixture` are boolean: they are *True* if, respectively, the suspect and the victim are contributors to the mixture. In the interaction with instance nodes **p1gt** and **p2gt**, respectively, they have the same function that node **p in mixture?** has, in the interaction with node **trace** of the class `Trace`.

The node **target** is the logical combination of the nodes **p1=s?** and **p2=v?** into four hypotheses:
h_1: Victim and suspect contributed to the mixture,

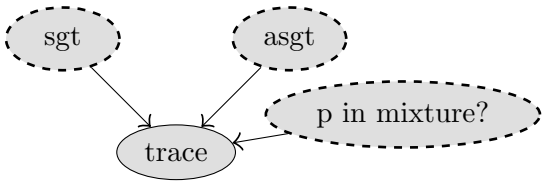


Figure 1.11: The class `Trace`. The output node **trace** represents the crime scene trace, and is identical to the nodes **sgt** (or **asgt**), depending on the state of the input node **p in mixture?**. For example, $\Pr(\textbf{trace} = \{a,b\} |\ \textbf{sgt} = \{a,b\}, \textbf{asgt} = \{c,d\}, \textbf{p in mixture?} = \textit{True}) = 1$, for any a, b, c, d. Nodes **sgt** and **asgt** are input nodes, bounded to nodes **gt** of the instances **s** and **as** of the class `Genotype`. Information about the genotype of the suspect is entered by fixing values of the node **gt** of the instance **s**.

h_2: Victim and an unknown person contributed to the mixture,
h_3: An unknown person and the suspect contributed to the mixture,
h_4: Two unknown persons contributed to the mixture.

However, by fixing the node **p2=v?** to its state *True*, the network can be used to evaluate the strength of the evidence with respect to the hypotheses h_1 and h_2, corresponding to h_p and h_d, respectively.

Node **mix** represents the alleles present at the specific marker of the mixed trace. It is the logical combination of the alleles from the two contributors. Information about the genotype of the suspect and the victim is entered by fixing the values for node **gt** in the instances **s** and **v** of the class `Genotype`. The likelihood ratio is then obtained by consecutively selecting the two hypotheses of interest (h_1 and h_2) in node **target** and reading the posterior probability corresponding to the observed state of node **mix**.

Again, this network models a single marker, but an OOBN to take into account results from all markers can be constructed, as shown in Figure 1.12. All markers (here only six for the sake of example) are modelled through instances **Marker i** ($i = 1, ..., 6$) of the network `Single Trace`. The input node **target** of each instances is bounded to the node **target** of the external class, as shown in Figure 1.12.

In order to support scientists in the evaluation of DNA mixture evidence from a quantitative point of view, object-oriented Bayesian network have been proposed in Cowell et al. (2007b, 2008), further developed in Cowell (2009), to take into account additional aspects, such as allelic drop-outs, stutter bands and silent alleles. Further uses of object-oriented Bayesian networks can be found in Cavallini and Corradi (2006), in the context of database searching problems (Gittelson et al., 2012) and in Green and Mortera (2009), where they are used to explore what happens when standard assumptions about the founder genes − such as the knowledge of the allele population proportion or the Hardy-Weinberg equilibrium − are violated.

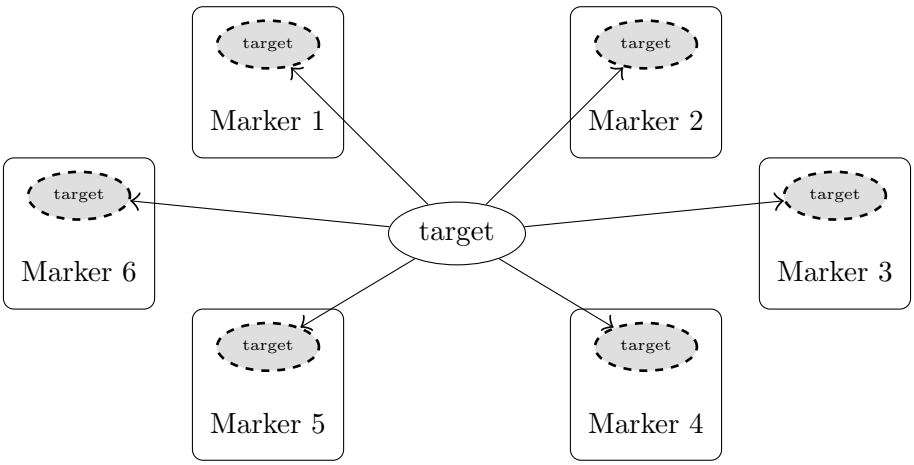


Figure 1.12: Expanded representation of an OOBN combining several markers.

1.4 Nonparametric Bayesian priors

A Bayesian nonparametric model is a Bayesian model on an infinite-dimensional parameter space. For our applications, the unknown parameter is typically represented by the vector containing the frequencies of some genetic characteristics (single STR alleles at one locus, or entire Y-STR haplotypes) in the population of possible perpetrators. The parameter space can be thought of as infinite dimensional, assuming that any STR number can potentially be observed. Even though this is not realistic (there may be biological reasons that bound the STR range) an infinite dimension can be a good approximation of a large dimension. This allows us to use nonparametric priors, in particular the two-parameter Poisson Dirichlet prior, which shows many interesting properties.

1.4.1 The two-parameter Poisson Dirichlet distribution

The two-parameter Poisson Dirichlet distribution, first introduced in Pitman (1992), is a distribution over ∇_∞, the infinite simplex of the form $\nabla_\infty = \{(p_1, p_2, ...) \mid p_1 \geq p_2 \geq ... > 0, \sum p_i = 1\}$. It is, in practice, a distribution on distributions. It can be constructed in two steps: first by simulating another distribution and then by sorting the results.

1. Given $0 \leq \alpha < 1$, and $\theta > -\alpha$, the vector $\mathbf{W} = (W_1, W_2, ...)$ is said to be distributed according to the GEM(α, θ), if $\forall i, W_i = V_i \prod_{j=1}^{i-1}(1-V_j)$, where the V_i are independently distributed as Beta$(1-\alpha, \theta + i\alpha)$. The GEM distribution (short for 'Griffin-Engen-McCloskey distribution') is well-known in literature as the "stick breaking prior", since it measures the random sizes in which a stick is broken iteratively.

2. The random vector $\mathbf{P} = (P_1, P_2, ...)$ obtained sorting the elements in $\mathbf{W}$ in decreasing order, has the two-parameter Poisson Dirichlet distribution PD(α, θ). Parameter α is called *discount parameter*, while θ is the *concentration parameter.*

For $\alpha = 0$, we obtain the well-known Poisson Dirichlet distribution, first introduced in Kingman (1975) as the infinite dimensional generalization of the classical Dirichlet distribution.

Notice that also $\alpha < 0$ and $\theta = -m\alpha$ for some $m \in \mathbb{N}$ is allowed: this is used for a model with only finitely (m) many DNA types, where a Dirichlet prior is put over the probability vector $\mathbf{P}$, with all hyperparameters equal to $-\alpha$. This case is not of interest for our research.

1.4.2 Pitman sampling formula

Partitions of the set $[n] = \{1, ..., n\}$ will be denoted with $\pi_{[n]}$, random partitions of $[n]$ will be denoted as $\Pi_{[n]}$. The different subsets forming a partitions are called 'blocks' of the partition.

Given a sequence of integer-valued random variables $X_1, ..., X_n$, consider the equivalence relation $i \sim j$ if and only if $X_i = X_j$. The equivalence classes of this relation form a random partition of $[n]$, which will be denoted as $\Pi_{[n]}(X_1, X_2, ..., X_n)$.

It holds that, if

$$\mathbf{P} \mid \alpha, \theta \sim PD(\alpha, \theta), \quad \text{and} \quad X_1, X_2, ... \mid \mathbf{P} = \mathbf{p} \sim_{\text{i.i.d}} \mathbf{p}, \tag{1.5}$$

then, for all $n \in \mathbb{N}$, the random partition $\Pi_{[n]} = \Pi_{[n]}(X_1, ..., X_n)$ has the following distribution:

$$\Pr(\Pi_{[n]} = \pi_{[n]} | \alpha, \theta) = \frac{[\theta + \alpha]_{k-1;\alpha}}{[\theta + 1]_{n-1;1}} \prod_{i=1}^{k} [1 - \alpha]_{n_i - 1;1}, \tag{1.6}$$

where n_i is the size of the ith block of $\pi_{[n]}$ (the blocks are here ordered according to the least element), and $\forall x, b \in \mathbb{R}, a \in \mathbb{N}$,

$$[x]_{a,b} := \begin{cases} \prod_{i=1}^{a-1} (x + ib) & \text{if } a \in \mathbb{N} \backslash \{0\} \\ 0 & \text{if } a = 0 \end{cases}.$$

This model can be used for partitions formed by a sample of individuals from a population with an infinite number of species, where the distribution of the vector containing the ordered population frequencies is assumed as $\text{PD}(\alpha, \theta)$ distributed. This is also known as the *Pitman sampling formula*, further studied in Pitman (1995). Notice that for $\alpha = 0$ we obtain the famous Ewens's sampling formula (Ewens, 1972).

1.4.3 The two-parameter Chinese restaurant process

Consider a restaurant with infinitely many tables, each one infinitely large. Let $Y_1, Y_2, ...$ be integer valued random variables that represent the seating plan: tables are ranked in order of occupancy, and $Y_i = j$ means that the ith customer seats at the jth table to be created. The two-parameter Chinese restaurant process is described by the following transition matrix:

$$Y_1 = 1,$$

$$\Pr(Y_{n+1} = i | Y_1, ..., Y_n) = \begin{cases} \dfrac{\theta + k\alpha}{n + \theta} & \text{if } i = k + 1 \\ \\ \dfrac{n_i - \alpha}{n + \theta} & \text{if } 1 \leq i \leq k \end{cases} \tag{1.7}$$

where k is the number of tables occupied by the first n customers, and n_i is the number of customers that occupy table i. The process depends on two parameters α and θ with the same conditions $0 \leq \alpha < 1$, $\theta > 0$ valid for the two-parameter Poisson Dirichlet distribution of Section 1.4.1. First described in case $\alpha = 0$ by Aldous (1985), this process is studied in details in Pitman and Picard (2006).

This process is deeply related to the two-parameter Poisson Dirichlet distribution thanks to the following results:

- if (1.5) and (1.7) hold, then for all $n \in \mathbb{N}$ the random partition $\Pi_{[n]} = \Pi_{[n]}(X_1, ..., X_n)$ has the same distribution as $\Pi_{[n]}(Y_1, ..., Y_n)$. They are both distributed according to the Pitman sampling formula (1.6). Stated otherwise, we can use the seating plan of n customers to obtain the same partition $\pi_{[n]}$ obtained by a sample $X_1, ..., X_n$ from a population whose probabilities are distributed according to a two-parameter Poisson Dirichlet distribution.
- the distribution of the infinite vector containing the relative sizes of the tables when infinite many customers have taken seats is $\text{PD}(\alpha, \theta)$.

This result is the key to use the two-parameter Poisson Dirichlet distribution as a Bayesian nonparametric prior for the rare type match case. Indeed, despite the rather complex definition presented in Section 1.4.1, it allows simplifying the definition of the only two probabilities of interest: given a database of size n, the probability (of interest for the prosecution) of observing a not yet observed Y-STR haplotype, and the probability (of interest for the defence) of observing the same not yet observed Y-STR haplotype twice. This will be explained in Section 2.7, and in details in Chapter 7, but the reader may already foresee that, by discarding enough information, the whole story can be described in terms of a customer entering into a restaurant with n customers already seated, and choosing an unoccupied table, or in terms of two customers entering the restaurant and choosing the same unoccupied table. These two probabilities can be easily obtained with formula (1.7).

1.4.4 The hyperparameters

To use the two-parameter Poisson Dirichlet prior, one has to deal with the presence of the two hyperparameters α and θ. This can be done either by choosing an hyperprior for them, or doing an assignment based on data (hence choosing an empirical Bayesian approach).

Regarding the possibility to infer their real values by analzying enough data, it is important to know that for a fixed $\alpha \in (0, 1)$, the $\text{PD}(\alpha, \theta)$ (for different θ) are all mutually absolutely continuous. Thus, θ cannot be consistently estimated. This is not much of a problem, since when n increases, the parameter θ becomes less and less important. It describes how much "social" are the customers: the smaller θ the more the customers tend to seat to already

occupied tables. It defines the size of the tables created first. On the other hand, α can be consistently estimated, as shown by the power law behavior, and there exists at least one consistent estimator for α (Carlton, 1999), namely:

$$\hat{\alpha} = \frac{\log K_n}{\log n}.$$

1.4.5 Power law behavior

Sampling from a two-parameter Poisson Dirichlet prior, using the marginalization described in (1.7), shows the rich-get richer clustering property (Teh et al., 2006): the more customers have been assigned to a table the more likely subsequent customers will be assigned to that table. Moreover, the more customers are assigned to unoccupied tables, the more the next customer will be assigned to an unoccupied table. The consequence of these two effects is that few tables will be occupied with many customers, and many tables will be occupied by very few customers. More precisely, the behaviour of the p_i follows the so-called *power law behaviour* (Newman, 2005), very common in many natural phenomenon, such as Y-STR data. This is one of the main reasons why we decided to use this prior for Y-STR data.

More precisely:

- Let K_n denote the random number of blocks of a partition $\Pi_{[n]}$ distributed according to the Pitman sampling formula with parameters α and θ. There exists a positive random variable S_α such that
$$\lim_{n\to+\infty} \frac{K_n}{n^\alpha} = S_\alpha, \quad \text{a.s.} \tag{1.8}$$
The distribution of S_α is a generalization of the Mittag Leffler distribution (Gorenflo et al., 2014).

- If $\mathbf{P} \sim \mathrm{PD}(\alpha, \theta)$, then
$$\frac{P_i}{Z i^{-1/\alpha}} \to 1, \quad \text{a.s., when } i \to +\infty \tag{1.9}$$
for a random variable Z such that $Z^{-\alpha} = \Gamma(1-\alpha)/S_\alpha$.

Chapter 2

Results

This Phd thesis consists of several challenging problems concerning statistical DNA evidence evaluation, investigated in a series of distinct studies, whose details are presented in Chapters 3 to 8. These studies, explained and summarized in the sections to come, are interrelated and follow a logical and chronological path.

2.1 Object-oriented Bayesian networks for evaluating DIP-STR profiling results from unbalanced DNA mixtures

This section is intended as a summary of the research contained in Cereda et al. (2014b), reproduced in full in Chapter 3.

DIP-STR markers, described in Section 1.1.5, were proposed as a novel type of genetic markers in Castella et al. (2013). The available kit for the DIP-STR analysis allows one to obtain the DIP-STR profile of a DNA mixture, to be compared with the DIP-STR profile of one or more questioned contributors, at a certain number of loci. This new set of markers is especially useful in case of extremely unbalanced mixtures, where the standard STR marker system fails to detect the alleles of the minor contributor. However, to exploit completely their potential, it is important to build an evaluative framework which helps in the assessment of observed profiling results in the light of the hypotheses of interest. This amounts to the calculation of the likelihood ratio, as described in Section 1.2.

To this extent, we decided to build an object-oriented Bayesian network to calculate the likelihood ratio for the two source level hypotheses 'the mixture contains the DNA of the victim and the suspect' (h_p), and 'the mixture contains the DNA of the victim and of an unknown person, unrelated to the suspect' (h_d). The data to evaluate is made of the DIP-STR results obtained from (i) DNA mixtures of two contributors, (ii) the known contributor (for instance, the victim of a sexual assault), (iii) the questioned contributor (the suspect). Figure 2.1 shows the OOBN whose details can be found in Chapter 3.

The probability tables of the class `DIP-STR` is filled with the so-called "Bayes estimates" of the allelic frequencies, based on a Dirichlet prior. They are obtained putting a Dirichlet prior

over the set of the allelic proportions, and updating this prior through the use of a database, seen as a multinomial observation. The posterior mean of each allelic proportion is used as estimate for the unknown allelic proportions. These estimates change from marker to marker, and from case to case. For this reason, an additional RHugin function is made available to automatise the procedure.

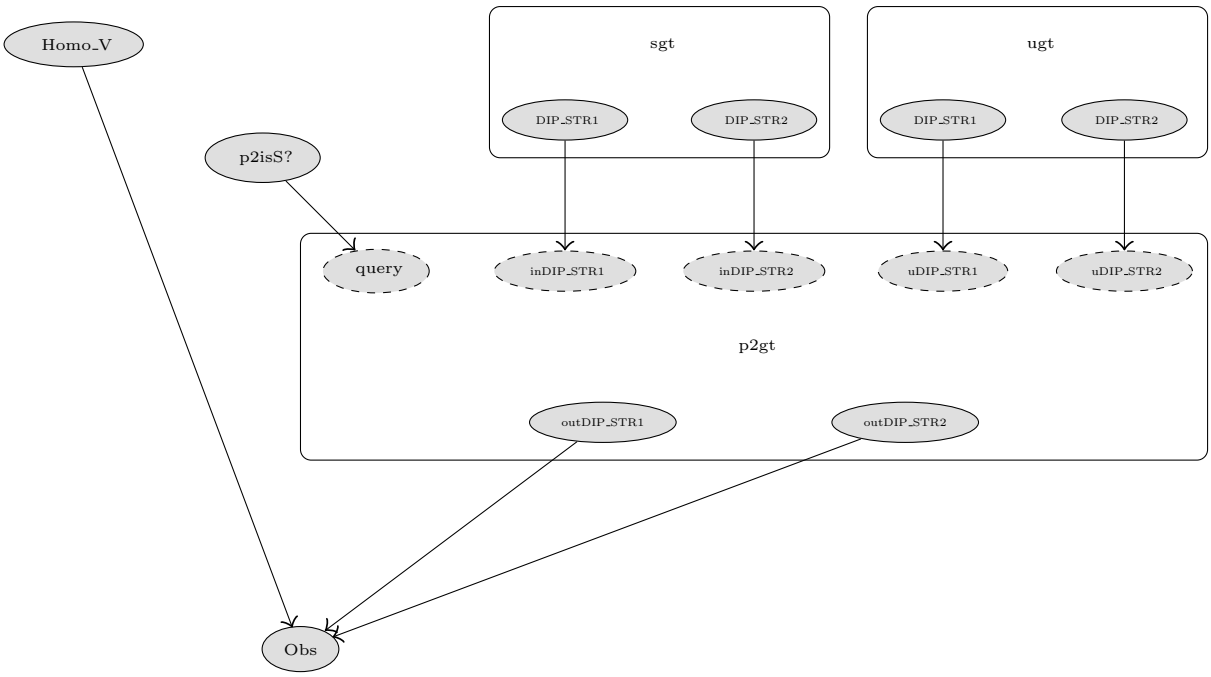


Figure 2.1: Expanded representation of the class `Marker`, modelling the mixture observation from a single locus. The victim is represented by the node **Homo_V**, with three states: *HomL*, *HomS* and *Hetero*, corresponding to his/her homozygosity or heterozygosity for the DIP allele. The right part of the network (i.e., all components other than **Homo_V** and **Obs**) models the minor contributor, that could be either the suspect, or an unknown person, whose alleles are represented by nodes **sgt** and **ugt**. The Boolean node **p2isS?** addresses the question of whether the second contributor is the suspect or an unknown person, while Node **p2gt** represents the genotype of the actual second contributor to the mixture. Node **Obs**, with states *La, Lb, Lab, Sa, Sb, Sab, X, nr*, represents the observed (minor contributor's) DIP-STR allele(s) in the trace. More details on the structure of each instance node, and on conditional probability tables can be found in Chapter 3.

Alternatively, several instances of the class `Marker` can be joined into the compound OOBN of Figure 2.2. The only modification required is to change node **p2isS?** inside the class `Marker` into an input node. Output node **H** is then linked to each node **p2isS?**.

Then, the compound likelihood ratio can be read as the ratio of the posterior probabilities of the state of node **H** when DIP-STR are entered, in virtue of the choice of equal prior probabilities over **H**.

A further step was that of modifying the interpretative model to be used when the suspect's DNA is not available for comparison, but that of a brother of the suspect is. The network `Marker for brother` riproduced in Figure 3.2 of Chapter 3, was built with that purpose.

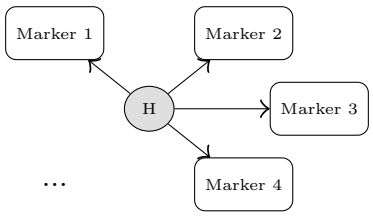


Figure 2.2: A compound object-oriented Bayesian network to be used for the evaluation of DIP-STR results from several markers. Each rectangular node is an instance of the class `Marker`, shown in Figure 2.1.

This is a good example of the flexibility of object-oriented Bayesian networks, which are readily adapted to different scenarios, in addition to provide a concise representation of the genotypic configuration of the various (assumed) contributors.

At the time when Cereda et al. (2014b) was written only 9 DIP-STR loci were available (Castella et al., 2013). However, the classes `Marker` and `Marker for brother` are usable with any marker. It is enough to adapt the meaning of the states and the probability tables of the input nodes. Hence, the same model can be used with the additional 9 loci that have recently been identified (Oldoni et al., 2015).

The paper contains also the application of the model to a casework example, where a relevant blood stain was collected on the body of a dead woman, and circumstantial evidence led to three suspects: a man and his two sons. The likelihood ratio for the hypotheses h_p and h_d defined above, was calculated marker by marker. Also, the likelihood ratio for the case in which the suspect's DIP-STR profile is not available, but that of a brother of the suspect is, is calculated, using the class `Marker for brother`. These examples allowed also to check that the likelihood ratio values obtained with the help of the probabilistic graphical model was the same as that obtained using a traditional formulaic approach. The advantage over the formulaic approach, which take different forms depending on the allelic configuration is clear, since except for the input values (i.e., the initial numerical specification), the structure of the OOBN remains constant. Moreover, formulae may become complicated when the scenario involves other members of the family of the suspect, depending on his degree of relatedness with the typed individuals. Thus, the use of an OOBN could also help to make evaluative procedures less prone to possible errors, since computations are entirely confined to the model. Note that the model presented in Figure 2.1 does not take into account extra variables of (potential) influence, such as typing errors and subpopulation effects.

2.2 An investigation of the potential of DIP-STR markers for DNA mixture analyses

This section is intended as a summary of the research contained in Cereda et al. (2014a), reproduced in full in Chapter 4.

The STR marker system is the most used set of markers for the analysis of DNA mixtures. It proved itself reliable for mixtures of any kind, except for extremely unbalanced mixtures,

for which it fails to detect the alleles of the minor contributor (see Section 1.1.5). This problem is usually overcome by the use of Y-STR markers, but they only work in case one contributor is female and the other is male. Also, they do not allow to distinguish between patrilineal relatives of the male contributor. The DIP-STR marker system hasn't got any of these limitations, and certainly can be preferred to STR markers and Y-STR markers in case of extremely unbalanced mixtures with a male major contributor (or a female minor contributor), or if it is necessary to distinguish between male relatives of the suspect, as in the casework example described in Section 2.1. However, an unconditional use of the DIP-STR technology, for any recovered trace, may not be the best option, given that available databases are sensibly smaller than the enormous amount of data already collected for STR and Y-STR. Also, the newness of this technology makes it more expensive. The question of which marker system is to be chosen is thus at the same time interesting and challenging.

In order to answer this interesting question, we decided to compare the DIP-STR marker system to the STR and Y-STR marker systems, from a statistical and forensic perspective. To do so, we contrasted the distribution of the likelihood ratio results obtained from 100,000 simulated cases using DIP-STR markers, with the distribution obtained (i) using traditional STR markers (assuming that we are in presence of moderately unbalanced mixture), and (ii) using Y-STR markers (assuming we are in presence of female-male mixtures). The comparison is performed under the prosecution's and the defence's case. For each marker system, each of the 100,000 simulated cases consists in the DNA profiles of the victim, of the suspect, and of another contributor. These are simulated by drawing locus by locus the alleles of the three individuals, independently and according to the allelic proportions in the population of interest. Under the prosecution's point of view, we "compose" a virtual mixture whose profile, at each locus, is made of the alleles of the victim and of the suspect. Under the defence's point of view, the profile contains alleles of the victim and of the other contributor. The virtual mixture obtained in this way is the one that could have been observed if we were in presence of real two-person mixtures, when the major contributor's genotype is available and under a set of assumptions, detailed in Chapter 4, concerning its quality.

For each simulated case, and each point of view, the likelihood ratio under the hypotheses "the mixture contains DNA material from the victim and the suspect" (h_p), "the mixture contains DNA material from the victim and an unknown contributor" (h_d) is calculated. All the likelihood ratios obtained for the prosecution's case are expected to be higher than 1, since we don't take into account the possibility of errors of laboratory. The higher the likelihood ratios obtained, the more the chosen marker system is interesting from the prosecution's point of view. On the other hand, in the defence's case, every time the genotype of the suspect is not compatible with the alleles in the mixture, a likelihood ratio of zero is obtained. The higher the number of zero values obtained, the more attractive is the chosen method from the defence's point of view.

In summary, the following conclusions can be made:

- For cases of, at worst, moderately unbalanced mixtures, the distributions of the likelihood ratio values both from the prosecution's and the defence's point of view suggest that the traditional STR marker system should be preferred.
- In case of extremely unbalanced mixtures STR markers are not reliable, but Y-STR

markers and DIP-STR markers can be used, the DIP-STR method should be preferred from the prosecution's point of view. From the the defence's point of view, preferences depend on whether one is more concerned with the strength of support obtained in case of a wrong association (i.e., when the likelihood ratio supports the wrong proposition), or on the number of times in which such a wrong indication is obtained.

The research went on with a discussion about the proportion of cases in which each of the three methods cannot be used: this has clearly a great influence on the choice of an analytical methodology.

The STR method is generally not useful to detect the minor contributor of extremely unbalanced mixtures, but in current practice, many (or most) extremely unbalanced mixtures probably go undetected (Oldoni et al., 2015), hence it is difficult to assess the proportion of cases in which such mixtures are encountered. In turn, it is easier to circumscribe the proportion of cases in which Y-STR markers are not usable: this happens whenever the major and the minor contributors are not female and male, respectively.

With regards to DIP-STR markers, the probability that an actual two-person mixture will not be recognised as such (i.e., the presence of a second contributor cannot be pointed out because the mayor is heterozygous or both contributors have the same DIP alleles) has been calculated. It turned out that about 4% of recovered stains, which are actually mixtures, will leave one with the uncertainty about the presence of a second contributor.

This allowed to conclude that the use of DIP-STR markers can be desirable for all those kind of traces that appear as a single stain with the use of STR and Y-STR markers, but for which one suspects the presence of a second contributor. In these cases, DIP-STR markers can also complement Y-STR results to discriminate paternally related individuals. Also, the use of DIP-STR markers could be of interest for all kinds of DNA stains, since one can establish in advance if they could be used, given that one starts by determining the DIP-STR genotype of the assumed known major contributor. In the case of a favourable outset, DIP-STR profiling can provide information about the second contributor in terms of one, two or no alleles. Although the likelihood ratio distributions obtained under the defence's and the prosecution's point of view are not as marked as those that can be obtained with traditional STR markers, they can still be regarded as practically useful. Moreover, new DIP-STR markers have been recently located (Oldoni et al., 2015). This will favourably improve the likelihood ratio distributions that could be obtained under the various competing points of view in a near future. However, analysts should also remind that the definition of the practical procedures will also encompass additional factors such as time and monetary constraints.

2.3 Some methodological issues

The research explained in the sections to come combines the attempt of finding a solution to the rare type match problem described in Section 1.2.4, and the treatment of some methodological issues encountered while working with DIP-STR data and, more generally, studying the state of the art of forensic DNA evaluation. There are four main methodological issues

discussed:

Plug-in likelihood ratio assessment. While building the interpretative framework for DIP-STR results, we were confronted with the necessity of quantifying DIP-STR allelic proportions having only a small database to represent the population of interest. A "full Bayesian" procedure would consist in choosing a prior distribution for this nuisance parameter, and integrating it out. However, in the researches summarized in Sections 2.1 and 2.2, we decided to follow the common forensic procedure, which estimates the allelic proportions using the posterior mean, after the observation of the database. In fact, this is only (at best) an approximation to the full Bayesian procedure. In Chapter 6 we call this approach 'hybrid' since it is reminiscent of the frequentist approach. The use of these plug-in methods, seen as an approximation of the exact Bayesian likelihood ratio, is not a problem in itself, and may present advantages in terms of computability, but the accuracy of these approximations should be carefully investigated. Moreover, we will show that often the full Bayesian procedure is not more difficult to use.

Evidence and background This issue is deeply related to the discussion about the use of hybrid plug-in methods described here above. In Section 1.2, the likelihood ratio is defined as the ratio of the probabilities of observing data D under the competing hypotheses of interest. It is important to discuss what to consider as D, since there are diverging options in literature. In our opinion, the data available to the expert can be divided into *evidence*

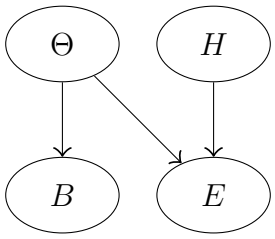


Figure 2.3: Bayesian network representing the dependency relations between E (evidence of the case), B (background data in the form of a database), Θ (population parameter), and H (hypotheses of interest).

and *background*. The first refers to data related to the crime, directly useful to discriminate between hypotheses of interest. The second refers to data which is not related to the crime, but is made available to the expert in order to help him to deal with the nuisance parameter(s) of the model. For instance, the evidence may consist of the DNA stain found at the crime scene, and of a suspect with the same DNA profile. The nuisance parameter of this model is θ, the population proportion of this DNA profile, and background data is a database from the population of possible perpetrators. Let us denote with E, B and Θ the random variables that correspond to evidence, background data, and nuisance parameter, respectively. The conditional dependency relation between these variables (together with H) are represented by the Bayesian network of Figure 2.3.

We believe that a Bayesian statistician or a Bayesian forensic scientist would use all available data to assign the value of the evidence. Hence, the Bayesian likelihood ratio should be

defined as

$$\mathrm{LR} = \frac{\Pr(E = e, B = b \mid H = h_p)}{\Pr(E = e, B = b \mid H = h_d)}. \tag{2.1}$$

This is in discrepancy with the definition of the likelihood ratio which can be found in much literature where only evidence data E is considered, with no mention for B.

It is true that B is independent of H, and that under the frequentist approach B and E are independent given H. This allows to simplify B from the formula 2.1. However, as can be deduced from the Bayesian network of Figure 2.3, this is not valid in a Bayesian framework, unless Θ is known.

Levels of uncertainty For a frequentist statistician, the likelihood ratio is a ratio of probabilities usually based on a model which is at best only a good approximation to the truth, and whose parameters have to be estimated using a limited sample. Thus, in a frequentist framework, along with the first basic initial uncertainty about the hypotheses, two more levels of uncertainty arise in the attempt of calculating the likelihood ratio. Some forensic literature (Morrison, 2010; Stoel and Sjerps, 2012; Curran et al., 2002; Curran, 2005) already pointed out the necessity for uncertainty assessment in the estimation of the likelihood ratio, even though they don't differentiate among levels.

On the other hand, for a true Bayesian individual these additional levels of uncertainty are part of the model, so that the definition of the Bayesian Pr includes not only beliefs about chances when picking people from the population, but also beliefs about parameters of the model, and beliefs about model. Hence, these levels could in principle be taken care of within the same framework.

There is a debate in literature (e.g. Taroni et al., 2016; Sjerps et al., 2016; Berger and Slooten, 2016; Curran, 2016) as to whether it makes sense to talk about 'estimation' and 'uncertainty assessment' regarding the likelihood ratio. Both the points of view are valid, depending on the context: if a frequentist approach to probability definition is chosen, it is pertinent to talk about estimations and uncertainty assessment. On the other hand, in a full Bayesian context, with Bayesian probabilities subjectively defined, they are misplaced. Moreover, most of the time, the Bayesian procedure consists of choosing priors (and hyperpriors) which are a compromise between personal beliefs and mathematical convenience. Additionally, Bayesian forensic statistics makes use of approximations, which may be seen as frequentist. It is thus interesting to investigate how good the choice of such priors is. Hence, whether Bayesian or frequentist approaches are chosen, the attempt at producing the likelihood ratio may lead to several levels of uncertainty which should be accounted for.

This subject matter is studied and discussed in the research summarized in Section 2.4.

Different data different likelihood ratios The evaluation of the totality of the data at the expert's disposal is often of difficult fulfilment. This is why often statisticians resort to reducing data to something less informative, but of more feasible evaluation. Especially in presence of many nuisance parameters, it can be wise to discard the part of data which primarily regards the nuisance parameters, and only indirectly regards which hypothesis is

more likely to be true. The more the data is reduced, the easier and more precisely the likelihood ratio for that reduction is estimated. However, each reduction comes with a cost: the stronger the reduction, the less the corresponding likelihood ratio value is helpful to discriminate between the two hypotheses. A compromise has to be made, between the gain in terms of precision and the loss in terms of strength of the evidence. Thus, one has to strike a balance between weakening the likelihood ratio and gaining in the precision of the estimate.

Many different methods to calculate the likelihood ratio are proposed in the literature. They are divided into Bayesian and frequentist, and most of the time they use different reductions of the data. In practice, the different methods are not suggesting different ways to obtain the same likelihood ratio. On the contrary, they are providing different methods to obtain different likelihood ratios: each reduction corresponds to a different likelihood ratio to be estimated. The choice of the reduction is often only implicit and one of the aim of this research is to make explicit the reduction corresponding to the proposed methods.

In the research described in the sections to come, several models for the likelihood ratio assessment in the rare type match case for Y-STR data are proposed, each using a different reduction of the data. The entirety of data to evaluate would be $D = (E, B)$, where E is composed by E_s (suspect's Y-STR hayplotype), E_t (crime stain's Y-STR haplotype, matching with the suspect), and B is a reference database of size n, containing a sample of n Y-STR haplotypes from the population of possible perpetrators. Each method corresponds to a different reduction of the data D, to be evaluated in the light of the two hypotheses h_p ('The crime stain was left by the suspect'), and h_d ('The crime stain was left by someone else').

2.4 Impact of model choice on LR assessment in case of rare haplotype match (frequentist approach)

This section is intended as a summary of the research published in Cereda (2016b), reproduced in full in Chapter 5.

Even though the likelihood ratio – seen as a way to update prior beliefs in the light of new observations – is motivated by Bayes' theorem, it may be of interest for frequentist statisticians as well, as a tool to measure the evidential value of data. With the aim of discussing the last two methodological aspects listed in Section 2.3, we studied and compared two frequentist methods to estimate the likelihood ratio in the rare type match case: the discrete Laplace method (see Section 1.2.5), and a modification of the nonparametric Good-Turing estimator (Good, 1953). Beside representing two interesting solutions to the rare type match problem with Y-STR data, they are also useful to show that:

(i) when the likelihood ratio is defined and estimated in a frequentist way, there are different levels of uncertainty that come into play, which have to be investigated and carefully discussed.

(ii) in a frequentist framework, the data to evaluate can be reduced in several ways, each

entailing the definition of a different likelihood ratio. The reduction is usually performed in order to reduce the error, since likelihood ratios for reduced data are more precisely estimated than likelihood ratios for more data, but it has to be clearly discussed, and justified. It is also important to understand that any reduction comes with a cost.

The discrete Laplace method First proposed in Andersen et al. (2013b), the discrete Laplace method is based on the model assumption that a single locus Y-STR allele is distributed according to the discrete Laplace distribution described in Section 1.2.5. The R package `disclapmix` (Andersen, 2013) allows one to estimate the frequency of any Y-STR haplotype, performing some statistical inference on a limited database. The method estimates the frequencies of unobserved haplotypes as well, hence we decided to apply it to the Y-STR rare type match problem. The data to evaluate is the Y-STR haplotype of the suspect and of the recovered stain, along with the list of Y-STR haplotypes in the available database. The data is not reduced, and the likelihood ratio that we want to estimate is equal to $1/f$, where f is the unknown frequency of the suspect's Y-STR haplotype in the population, to be estimated with the use of the `disclapmix` package. The reciprocal of this estimate is used as an estimate of the likelihood ratio.

The generalized Good method The Good-Turing estimator, first described in the famous paper Good (1953), is a nonparametric nearly unbiased estimator for the total probability of the species which are unseen in a sample of size n, obtained by drawing species independently from the population. This estimator is equal to $\frac{N_1}{n}$, where N_1 is the number of singletons in the sample. We decided to build an estimator of the likelihood ratio for the rare type match problem, based on the Good-Turing estimator. Indeed, by reducing the data to the simple fact that the Y-STR haplotype of the crime stain matches the Y-STR haplotype of the suspect, but they are not in the database (hence, discarding information about the specific Y-STR haplotype of the suspect, and those in the database), the numerator of the likelihood ratio is precisely the probability of observing an unseen species at the $n+1$st observation. The denominator is the probability of observing the same unseen species twice (both in the $n+1$st and $n+2$nd observations). We proved that, as N_1/n is nearly unbiased for the numerator, $\frac{2N_2}{n(n-1)}$ is nearly unbiased for the denominator (at least when n is big enough). Thus, $\frac{N_1 n}{2N_2}$ can be used as estimate for the likelihood ratio. We call this estimator the "generalized Good estimator".

These two methods are a good example of different reductions of the data. The discrete Laplace method allows one to estimate a likelihood ratio that evaluates almost[1] the entirety of the data at disposal, while the generalized Good method is suitable to estimate likelihood ratios that only evaluate part of the data: information about the particular Y-STR haplotypes of the suspect and those in the database are discarded. Hence, the likelihood ratio values obtained with the generalized Good method were expected to be smaller than those obtained with the discrete Laplace method. Thus, before directly comparing the likelihood ratio values obtained with the two estimators, we compared them with the values they aim at estimating.

[1] the database is reduced to count so we loose the information about the order in which data is listed.

Moreover, we discussed and quantified the uncertainty that is involved in each of the two estimates. To do so, for each method we simulated many cases, for which we were able to calculate the true likelihood ratio and the estimates, and we studied the difference of their logarithm. Notice that to actually know the true likelihood ratio it would be necessary to know the entire population. However, the available Y-STR database contains approximately 19,000 haplotypes from 129 different locations in the world (Purps et al., 2014). We consider only 7 loci out of the 23 available, and we pretend that the database contains the entire population of interest for our case.

In the discrete Laplace case, the error is both due to the fact that a specific distribution was chosen to model the Y-STR alleles data (the discrete Laplace), and to the fact that parameters for that distributions are estimated using databases of limited size. For the generalized Good method the entire error is due to the approximation step.

Comparing the distributions of the error is not enough. Indeed, one has to take into account the fact that the generalized Good method discards a lot of data and thus is less useful for the final aim of the evaluation (being able to distinguish the correct hypothesis). One can say that the hallmark desired likelihood ratio is one that evaluates the entirety of the data at disposal, thus $1/f$, hence a comparison with it is also proposed (see Chapter 4, for the details).

2.5 A useful Lemma

In a forensic casework, it is very common that prosecution and defence agree on part of the available information, and disagree on other. For instance, the prosecution may consider the correspondence between the profiles of two DNA traces a sure event (since they came from the same source), while the defence may believe that this correspondence is due to coincidence. Nevertheless, both parts may accept a database as representative of the population of possible perpetrators. Mathematically, one can say that prosecution and defence disagree on the distribution of the evidence, while they agree on the distribution of the background data. This situation is represented by the Bayesian network of Figure 2.4, where node Y stands for the entirety of data at disposal (say the DNA profile of the crime stain, of the suspect and a list of DNA profiles in the form of a database), while node X is the part of data whose distribution is agreed on by prosecution and defence (only the suspect DNA profile and the database). The distributions of X and Y depend on the nuisance parameter(s) represented by A (such as the vector containing the population proportions of all DNA profiles in the population of possible perpetrators).

The Lemma is very useful to simplify the development of the likelihood ratio for the evaluation of X and Y in this common forensic situation, since it is enough to calculate two posterior expectations. In most of the cases of interest X represents everything except the observation of the crime stain (hence, the reference database and the suspect's DNA profile). The probability of the entirety of data given X and the parameter(s) is 1 according to the prosecution, while is a function of the parameter according to the defence.

Lemma. *Given four random variables A, H, X and Y, whose conditional dependencies are represented by the Bayesian network of Figure 2.4, the likelihood function for h, given $X = x$ and $Y = y$ satisfies*

$$\text{lik}(h \mid x, y) \ \propto \mathbb{E}(p(y \mid x, A, h) \mid X = x).$$

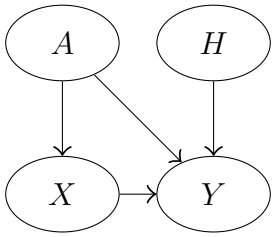


Figure 2.4: Conditional dependencies of the random variables of the Lemma.

Hence the likelihood ratio can be developed as

$$\text{LR} = \frac{p(x, y \mid h_p)}{p(x, y \mid h_d)} = \frac{\mathbb{E}(p(y \mid x, A, h_p) \mid X = x)}{\mathbb{E}(p(y \mid x, A, h_p) \mid X = x)}.$$

This Lemma is used for the development of complex likelihood ratio formulae in Cereda (2016a), Cereda et al. (2016), and Cereda (2016c), where a proof is also given. These publication are reported in full length in Chapters 7, 2.8, and 2.7, respectively.

2.6 Bayesian approach to LR for the rare match problem

This section is intended as a summary of the research published in Cereda (2016a), reproduced in full in Chapter 7.

The first two methodological issues detailed in Section 2.3, prompted us to study the two most common Bayesian solutions used in forensic science, the beta-binomial and Dirichlet-multinomial model. Again, we have a double aim: to customise these models and make them suitable as solution for the rare type match problem, and to show the difference between the widespread plug-in solutions and a proper full Bayesian approach to likelihood ratio assessment, which is obtained by evaluating also background data.

The beta-binomial model for the rare Y-STR haplotype match problem If the data to be evaluated is made of the suspect's and the crime stain's Y-STR haplotypes (E_s and E_c) which correspond, and of a database containing n Y-STR haplotypes from the population of possible perpetrators, the beta-binomial model can be adopted. This amounts to put a beta prior over θ (the unknown population proportion of the suspect's Y-STR haplotype) and to represent the database as a binomial variable B with parameters n and θ. The use of a full Bayesian approach which evaluates both $E = (E_s, E_c)$ and B leads to the following likelihood ratio (details can be found in Chapter 6):

$$\text{LR} = \frac{\alpha + \beta + n + 1}{\alpha + b + 1}. \tag{2.2}$$

where α and β are the shape parameters of the beta prior, and b is the number of times the suspect's Y-STR haplotype is observed in the database.

On the other hand, the plug-in approach would lead to the following estimate for the likelihood ratio:

$$\widehat{\mathrm{LR}} = \frac{\alpha + \beta + n}{\alpha + b}.$$

Sensitivity analysis of the logarithms of these two quantities, when $b = 0$ (i.e., in the case of a rare type match problem) and of their difference, shows that the two quantities depend a lot on α while they do not depend much on β. On the whole, the plug-in estimate $\widehat{\mathrm{LR}}$ is more sensitive than LR to changes in α, and is more anti-conservative (it always exceeds the full Bayesian likelihood ratio). The difference, however, is important only for small values of α. Otherwise, the two methods would lead essentially to the same conclusions, so that the plug-in can be seen as a good approximation of the proper Bayesian procedure.

The Dirichlet-multinomial model for the rare Y-STR haplotype match problem With the same data to evaluate, another possibility is to treat the database as a multinomial sample of size n. For a population with k different Y-STR haplotypes, the conventional choice for the prior over the parameter vector $\boldsymbol{\theta}$, containing the population frequencies of all the different haplotypes in Nature, is the Dirichlet distribution with all the hyperparameters equal to the same value α. The reason for that, is that our prior knowledge about the probabilities of the different categories is invariant to permutations of the categories. Green and Mortera (2009) proposed a similar model, along with an implementation, using an equivalence with a Polya urn scheme to avoid dealing with continuous parameter θ. Their implementation, though, is not directly exploitable for our model, since they assume that k is known. For several applications, k is chosen equal to the number of types in the database, but this is not appropriate for Y-STR haplotypes, since the database usually does not offer a good coverage. We don't even have an idea of a maximal number of categories, so we proposed a modification of the classical Dirichlet-multinomial model, where k is now modelled as a random variable, with a prior $p(k)$.

The full Bayesian likelihood ratio for this model is developed (for α=1) and is equal to

$$\mathrm{LR} = \frac{1}{2}\frac{\displaystyle\sum_{k=k_{obs}+1}^{m} \binom{k}{k_{obs}+1} \frac{\Gamma(k)}{\Gamma(k+N+1)} p(k)}{\displaystyle\sum_{k=k_{obs}+1}^{m} \binom{k}{k_{obs}+1} \frac{\Gamma(k)}{\Gamma(k+N+2)} p(k)}, \tag{2.3}$$

where k_{obs} is the number of distinct Y-STR haplotypes observed in the database. In our research two prior distributions over k are studied: the Poisson distribution and the negative binomial distribution. The full Bayesian likelihood ratio obtained with these priors can be compared to likelihood ratio estimated through the classical Bayesian plug-in method. This, for $\alpha = 1$, is equal to

$$\widehat{\mathrm{LR}} = \bar{k} + N, \tag{2.4}$$

where the number of haplotypes is a fixed value $\bar{k}$, to be chosen (or estimated) in advance. Both using a Poisson prior and a negative binomial prior over k, we decided to perform a

sensitivity analysis of the logarithms of the two likelihood ratios (2.4) and (2.3), and of their difference, when $\bar{k}$ is chosen equal to $\mathbf{E}(K)$.

Results show that, using the Poisson prior, the plug-in approach can be seen as a rather good approximation of the full Bayesian likelihood ratio, while with the negative binomial prior over K the difference between the two can be quite significant, especially if the mean value for K is big. Moreover, for both the priors, the plug-in and the full Bayesian likelihood ratios strongly depend on the hyperparameters, and not much on the data. In particular, they depend on the data only through k_{obs}, the number of distinct observed types, and not on their frequencies. On the other hand, both these quantities depend much on the mean value of K.

This research pointed out the inadequacy of both models (beta-binomial, and Dirichlet-multinomial) for the rare type match problem. Indeed, for both models, the prior over the parameter is chosen for mathematical convenience, rather than because it does represent the expert's belief. This procedure would be sensible only if, at the end, the data overrules the prior choice, whilst this is not the case. This is the reason why we decided to investigate different priors, more realistic for Y-STR data, such as those proposed in the study described in Chapter 7.

2.7 Nonparametric Bayesian approach to LR assessment in case of rare haplotype match

This section is intended as a summary of the research published in Cereda (2016c), reproduced in full in Chapter 7.

The need of priors which better represent Y-STR haplotypes frequencies, prompted us to study new kinds of distributions. In particular, we realized that one of the problems of the method used in Cereda (2016a) (see Section 2.6), is the choice of equal hyperparameters for the Dirichlet prior, since it generates a posterior distribution for θ (after the observation of the multinomial sample) such that the probability of the non observed types is almost uniformly distributed over these types. In fact, looking at the distribution of the frequencies of thousands of Y-STR haplotypes collected all over the world (shown in Figure 2.5), one can realize that there are many types with very rapidly decreasing probabilities, showing a sort of power-law behavior, described in Section 1.4.5. A tempting solution would be to use an asymmetric Dirichlet prior. However, this would lead to extremely difficult computations. The density of the two-parameter Poisson Dirichlet distribution (described in Section 1.4.3) shows a behavior similar to that of Figure 2.5, at least for the smaller probabilities (those of interest in the rare type match case). For this reason the use of this distribution as a prior over the nuisance parameter was investigated. The necessary assumption for this solution is that there are infinitely many Y-STR haplotypes in nature. This assumption, already used by Kimura (1964), is acceptable given the huge number (almost four thousand) of distinct 7 loci Y-STR haplotypes observed in the available database. To use this model, the additional assumption that the specific sequence of STR alleles forming each Y-STR haplotype carries no information was made. This is not realistic, since the distance of Y-STR numbers between

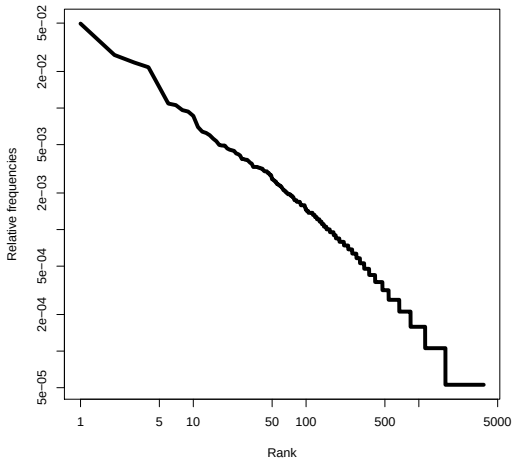


Figure 2.5: Power-law behaviour (logarithmic scale) observed for the ranked frequencies of 7 loci Y-STR haplotypes of Purps et al. (2014).

two haplotypes is an indication of the number of mutation drifts that separate them, but we considered this information too complex to be exploited. Thanks to this assumption, one could reduce by sufficiency the database of size n to a partition of the set $[n]$, obtained by grouping in the same subsets the indexes corresponding to the same haplotypes. According to this model, prosecution and defence agree that this partition is distributed according Pitman's sampling formula (see Section 1.4.2). The entirety of data is made of $n+2$ observed hyplotypes (those in the database and two additional haplotypes from the suspect and from the crime scene trace), and can be reduced to a partition of the set $[n+2]$, where $n+1$ and $n+2$ form a subset by themselves in the rare type match case. Prosecution and defence disagree on the distribution of this random partition since, according to prosecution, elements $n+1$ and $n+2$ are in the same class with probability one.

Using the Chinese Restaurant representation described in Section 1.4.3, and the Lemma presented in Section 2.5, and modelling the hyperparameters α and θ as random variables A and Θ, the likelihood ratio can be written as

$$\mathrm{LR} = \frac{1}{\mathbb{E}(\frac{1-A}{n+1+\Theta} \mid \Pi_{[n+1]} = \pi_{[n+1]})},$$

where $\pi_{[n+1]}$ is the partition obtained enlarging the database with the Y-STR haplotype of the suspect.

Empirical validation showed that choosing a flat hyperprior for the parameters α and θ, the likelihood ratio can be accurately approximated by

$$\mathrm{LR} \approx \frac{n+1+\theta_{MLE}}{1-\alpha_{MLE}}, \tag{2.5}$$

where α_{MLE} and θ_{MLE} are the maximum likelihood estimates of α and θ, respectively. This is equivalent to a plug-in approach where the parameters are estimated through data, and where estimates were plugged into the likelihood ratio. Here this approach is validated by

empirical studies, whose details can be found in Chapter 7. This result was unexpected, but very useful, since it makes the method very practical to be used, despite the complex theoretical background. The paper offers also a comparison between the values obtained with (2.5), and the 'true' likelihood ratio one would obtain knowing the vector $\mathbf{p}$ with the sorted population frequencies of all Y-STR haplotypes in Nature, both using the same reduction of data, and not reducing the data at all. To make this comparison, we built a Metropolis-Hastings algorithm similar to the one proposed in Anevski et al. (2013). This comparison led to the conclusion that the use of the two-parameter Poisson Dirichlet prior is very convenient, and allows one to obtain accurate approximation of the likelihood ratio one would obtain knowing the nuisance parameter $\mathbf{p}$. Additional details can be found in Chapter 7.

2.8 A solution for the rare type match problem when using the DIP-STR marker system

This section is intended as a summary of the research published in Cereda et al. (2016), reproduced in full in Chapter 8.

The available database of DIP-STR alleles contains only observations from about 100 Swiss individuals. Hence, whenever a new trace is analysed, it is very likely to observe alleles not contained in the database. Not all the solutions proposed and summarized in the previous sections for the Y-STR rare type match problem are appropriate for the DIP-STR case. For instance, the generalized Good method requires a large sample size, while the discrete Laplace and two-parameter Poisson Dirichlet model requires trust in the prior: one would need a sort of empirical validation that cannot be achieved given the small data available. On the other hand, the method proposed in Cereda (2016a), which uses a Dirichlet prior with a random number of categories can be successfully applied, and the model can be developed in a way that does not require the ad hoc plug-in approximations used in Cereda et al. (2014b) for the DIP-STR allelic proportions, and allows one to obtain the full Bayesian likelihood ratio.

The Bayesian network of Figure 2.6, is developed starting from the structure of the object-oriented Bayesian network presented in Cereda et al. (2014b), with two additional nodes, representing the database ($\mathbf{D}$) and the nuisance parameter ($\mathbf{\Theta}$).

The data to evaluate is made of a database containing n DIP-STR alleles from a relevant population, along with the two DIP-STR alleles of the victim, the two DIP-STR alleles of the suspect, and one or two alleles observed from the mixed trace. The hypotheses of interest are the same as those in Section 2.1 regarding whether the second contributor of the stain is the suspect or an unknown (unrelated) individual from the relevant population. We assume that there may be at most $2m$ possible DIP-STR alleles, but only some of them are actually present in the population and even less are those observed in the available database. The nuisance parameter $\boldsymbol{\theta} = (\theta_1^L, ..., \theta_m^L, \theta_1^S, ..., \theta_m^S)$, contains the population proportions of the $2m$ potential DIP-STR alleles.[2] The values corresponding to alleles which are not present

[2] L and S represent "long" and "short" as defined in Section 1.1.5.

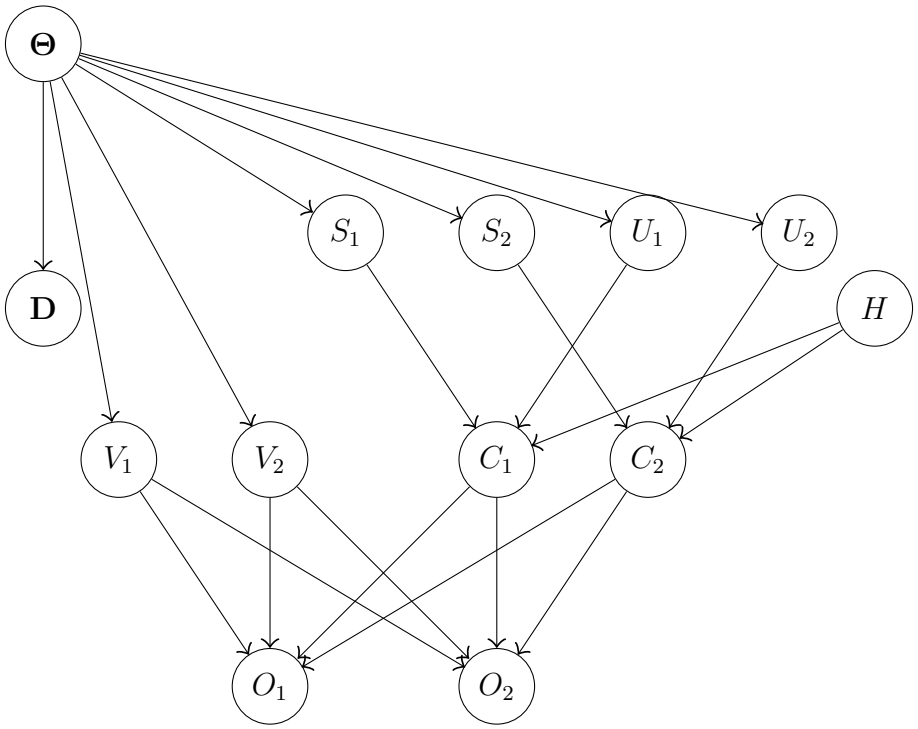


Figure 2.6: Bayesian network for the Dirichlet-multinomial model with a random number of types, to be used for single loci DIP-STR results from mixtures of two contributors. V represents the victim's (or the major contributor's) DIP-homozigozity at the represented locus, with three possible states: $HomoL$, $HomoS$ and $Hetero$. Node H represents the hypotheses h_p and h_d. Nodes S_1 and S_2 represent the two DIP-STR alleles of the suspect, while nodes U_1 and U_2 represent the two DIP-STR alleles of the alternative (unknown) minor contributor. Nodes C_1 and C_2 represent the DIP-STR alleles of the actual minor contributor. Depending on H they can be a copy of S_1 and S_2 (under h_p), or of U_1 and U_2 (under h_d). Nodes O_1 and O_2 contain the DIP-STR alleles observed from the mixed trace. Nodes S_1, S_2, U_1, U_2, C_1, C_2, O_1, and O_2 have states (L,i) (or (S,i)) where $i \in \{1, ..., m\}$ represents the STR part of the DIP-STR allele. Node $\boldsymbol{\theta}$ contains the population proportions of all the potential $2m$ DIP-STR alleles at that locus, (for instance, alphabetically ordered). For more details on the conditional probability distribution of each node, see Chapter 8.

in the population are zero. The prior for $\boldsymbol{\theta}$ is articulated and can be described as follows. The random variable ψ representing the sum of the frequencies of the DIP-STR alleles of type L has a uniform prior, while the normalized vector containing the frequencies of the DIP-STR alleles of type L (obtained dividing each θ_i^L by ψ) has a Dirichlet distribution with all hyperparameters equal to α, with a random number of categories (k^L). The same holds for the normalized vector containing the frequencies of the DIP-STR alleles of type S. k^L and k^S have uniform priors as well.

The Lemma discussed in Section 2.5, and proved in Cereda (2016c), is used to simplify the development of the full Bayesian likelihood ratio. Details can be found in Chapter 8.

The sensitivity analysis conducted on the likelihood ratio values, using data from different DIP-STR markers and different contributors, shows that these values moderately depend on α, at least when a uniform prior is given to k^L, k^S, and ψ. This shows the need of introducing better priors, either less sensitive to changes in α or more realistic, such as the prior used for Y-STR frequencies in Cereda (2016c) (see Section 2.7). Moreover, the full Bayesian likelihood ratio values are compared to those obtained with the plug-in approximation adopted in Cereda et al. (2014b). The values are practically identical.

The drawback of this model is, again, the fact that few information from the database are used in the likelihood ratio. This can be due to the choice of having all hyperparameters equal to α. In order to compensate for this undesired feature, another solution can be adopted. It consists of an hybrid combination with the Good-Turing estimator. Given k^L, and the observation from the database augmented with suspect's and victim's alleles, the posterior expectation of the total probability of the unobserved DIP-STR alleles of type L, is equal to

$$\frac{(k^L - k^L_{\text{obs}})\alpha}{k^L\alpha + n^L}, \tag{2.6}$$

where k^L_{obs} is the number of dinstinct DIP-STR alleles of type L in the augmented database. According to the Good-Turing estimator, the total probability of the unobserved DIP-STR alleles of type L can be estimated by

$$\frac{N_1^L}{n^L}, \tag{2.7}$$

where N_1^L is the number of singletons of type L, while n^L is the total number of alleles of type L in the augmented database. By equating (2.6) and (2.7), we obtain an empirical Bayes estimate of k^L, of the form

$$\hat{k^L} = \frac{N_1^L n^L + k^L_{\text{obs}}\alpha n^L}{\alpha n^L - \alpha N_1^L}.$$

This methods uses more information from the augmented database (namely n_1^L and n_1^S). Additional investigations led to the conclusion that it is a very good approximation to the full Bayesian likelihood ratio.

Part II

Papers

Chapter 3

Object-oriented Bayesian networks for evaluating DIP-STR profiling results from unbalanced DNA mixtures

This chapter is based on:
Cereda, G., Biedermann, A., Hall, D., and Taroni, F. (2014). Object-oriented Bayesian networks for evaluating DIP-STR profiling results from unbalanced DNA mixtures. *Forensic Science International: Genetics*, 8:159 - 169.

Abstract

The genetic characterization of unbalanced mixed stains remains an important area where improvement is imperative. In fact, with current methods for DNA analysis (Polymerase Chain Reaction with the SGM Plus^{TM} multiplex kit), it is generally not possible to obtain a conventional autosomal DNA profile of the minor contributor if the ratio between the two contributors in a mixture is smaller than 1:10. This is a consequence of the fact that the major contributor's profile 'masks' that of the minor contributor. Besides known remedies to this problem, such as Y-STR analysis, a new compound genetic marker that consists of a Deletion/Insertion Polymorphism (DIP), linked to a Short Tandem Repeat (STR) polymorphism, has recently been developed and proposed elsewhere in literature (Castella et al., 2013). The present paper reports on the derivation of an approach for the probabilistic evaluation of DIP-STR profiling results obtained from unbalanced DNA mixtures. The procedure is based on object-oriented Bayesian networks (OOBNs) and uses the likelihood ratio as an expression of the probative value. OOBNs are retained in this paper because they allow one to provide a clear description of the genotypic configuration observed for the mixed stain as well as for the various potential contributors (e.g., victim and suspect). These models also allow one to depict the assumed relevance relationships and perform the necessary probabilistic computations.

3.1 Introduction

The common way to analyze DNA mixtures for forensic purposes is to use the Polymerase Chain Reaction (PCR) and STR markers (Butler, 2011). This has proven to be a very successful technique, both for its speed and its high discriminating power. But besides its many advantages, this technique has also some drawbacks. When dealing with mixtures of two contributors, for example, the method will generally not work successfully if the proportion between the DNA of the two contributors is more extreme than 1:10 (Clayton and Buckleton, 2005).[1] These situations are quite common, such as in cases of sexual assaults when the victim's DNA is largely predominant, or in case of microchimerism during pregnancy or following organ transplant. To address this constraint, an alternative analytical method has recently been developed and proposed (Castella et al., 2013). This method is based on the use of a new compound marker, formed by an STR marker coupled to a marker in which a Deletion/Insertion Polymorphism (DIP) is known to be present.

DIPs as such have previously been discussed in biological and biostatistical literature (e.g., identification and characterization of di-allelic polymorphisms and allelic frequencies in particular ethnies and in natural population (e.g., Weber et al., 2002; Vali et al., 2008; Neuvonen et al., 2012)), genetics (e.g., identification of DIPs as causes of genetic diseases (e.g., Cooper and Krawczak, 1991)), and forensic science (e.g., use of DIPs for analysing highly degraded DNA (e.g., da Costa Francez et al., 2012)). The novelty of the paper here is to present an interpretative model that represents an essential element for rendering the potential of a new compound marker formed by a DIP marker coupled to an STR marker operationally useful for practitioners. The discussion will mainly concentrate on the coherent combination of the advantages of the two kinds of polymorphism, and on how this may be formally achieved through an interpretative model. In particular, this paper aims to develop and describe a probabilistic framework for the assessment of profiling results obtained with this novel typing technique, applied in the particular context of unbalanced DNA mixtures of two contributors. The approach relies on probabilistic graphical models, in particular object-oriented Bayesian networks (OOBNs). The paper also includes a discussion of this framework for two casework examples.

Section 3.2 provides a short description of the DIP-STR method from a biological point of view, while Section 3.3 describes the generic structure of the probabilistic model (i.e., OOBN) that has been built to evaluate DIP-STR profiling results. More detailed descriptions of the different structures composing the proposed OOBN are confined to Appendix A. Section 3.4 presents two casework examples to illustrate the kind of calculations that can be performed with the proposed graphical network environment (i.e., to obtain likelihood ratios for particular DIP-STR profiling results).They also exemplifiy the flexibility of graphical models, which are readily adapted to different scenarios. The last section presents a discussion and conclusions.

[1]Here, the threshold of 10% is retained as the limit of detection of the minor DNA for blood:blood mixtures. This value varies depending on the types of biological fluids which constitute the mixture and the specific combination of genotypes present in the mixture (as reported in (Applied Biosystems, 2012)) and should be assessed in the validation procedure (Butler, 2011).

3.2 Genetic background

The standard method for the analysis of DNA mixtures relies on STR primers as part of a procedure that seeks to amplify only selected portions of DNA, that is regions where particular STR markers are located. STR primers are only locus-specific, not allele-specific. This means that, as the DNA of both contributors have the same loci, these primers should, in theory, anneal to both the markers of the major and to those of the minor contributor. This is, in fact, what happens whenever the minor contributor's DNA represents more than (about) the 10% of the major contributor's DNA. But, below this threshold, the minor contributor's DNA is generally not detected, as it is "masked" by the DNA of the major contributor. The difficulties, in this case, include the detection threshold of most capillary electrophoresis equipments, possible amplification biases and low template amplification conditions for the minor contributor's DNA. As a result, its signal is lost under major alleles, stutters and background noise with the consequent failure in retrieving important information.

This problem can be addressed with the use of primers that are allele-specific, to assure that − each time the two contributors have different genotypes in some marker − the primers will anneal to different alleles. This thus would avoid situations that involve competition. Based on these considerations, DIP-STRs were recently proposed as novel type of genetic marker (Castella et al., 2013). The novelty consists on pairing a Deletion/Insertion Polymorphism (DIP) (e.g., Weber et al., 2002) with a standard STR, to form a superlocus where the two component loci are not independent (less than 500bp apart).[2] In this way, it is possible to design two alternative allele-specific primers overlapping the DIP locus, denoted L-DIP primer and S-DIP primer. Each of these is to be used together with a primer downstream the STR region.

Hence, DIP-STR genotyping allows the selected amplification of the minor contributor's genotype (DIP-STR genotypes of minor contributors were successfully typed at ratios as low as 1:1000), as long as it has alleles that are absent in the major contributor's genotype. The best scenario is when the DNA of, respectively, the major and minor contributor are homozygous for different DIP alleles (i.e., one S-S and the other L-L). In this case, the possible results can show either two different minor DNA haplotypes or one, depending on the STR-homozygosis or heterozygosis of the minor contributor. On the other hand, when the major contributor's DNA is DIP-homozygous and the minor contributor's DNA is DIP-heterozygous, only one haplotype of the minor DNA can be retrieved (i.e., the one concerning the DIP allele opposite to the DIP allele of the major contributor's DNA).

A limitation of this method is that, when the predominant DNA is DIP-heterozygous or both contributors are DIP-homozygous of the same type, it is not possible to have any information about the minor contributor's genotype, because both DIP primers (S and L), if used, will anneal to the major contributor's DNA. For such cases, the term *uninformative genotype* is used here. Table 3.1 summarizes the possible outcomes.

As a side note, it is worth mentioning that there is a traditional way to overcome the problem of strongly unbalanced mixtures in some cases. The use of Y-STR markers, for example, is of

[2]The two composing loci are not independent because they are so close on the chromosomes that they cannot recombine.

DIP genotype of major/minor contributor	**Number of haplotypes retrieved from minor contributor's DNA**	**Informativeness of genotypic configuration**
Hom/Hom (different allele)	2 (if STR het)	Yes completely
	1 (if STR hom)	Yes
Hom/Het	1 (regardless STR)	Yes
Hom/Hom (same allele)	0 (regardless STR)	No
Het/Hom	0 (regardless STR)	No
Het/Het	0 (regardless STR)	No

Table 3.1: Informativeness of genotypic configurations. 'Hom' denotes homozygous and 'Het' heterozygous.

great help for cases where a male component is detected in DNA mixtures with a high female background (Roewer, 2009). However, Y haplotypes can be quite common in a population (Vermeulen et al., 2009) and, if no mutations occur, patrilineal relatives of a suspect cannot be excluded as being the contributors to the stain. Recently, a panel of 13 rapidly mutating (RM) Y-STR markers has been identified (Ballantyne et al., 2012), which successfully differentiates between closely and distantly related males. However, both the classical and the RM Y-STR techniques are useful only for a specific sex mismatch, that is if the major contributor is a women and the minor contributor is a man.

One of the advantages of the DIP-STR method over the classic STR method is that, whenever it is feasible, it detects alleles that can directly be related to the second contributor. Conversely, with the classical STR method used for mixtures of two contributors, if in some locus less than four different alleles are present, it is impossible (unless the height of the peak is taken into account and this information is reliable) to discern the alleles that belong to the second contributor, despite knowing the genotype of the first. It can only be assessed that, if the second contributor is heterozygous, he shares one allele with the first contributor, but not to decide which one. With this new method, in a case of completely informative genotypic configurations (see first row of Table 3.1), the complete genotype of the minor contributor can be obtained, even if this individual shares some STR alleles with the main contributor. In addition, even in case of only partially informative configurations − in which only one allele is observed (see the second and third row of Table 3.1) − it is certain that the detected allele belongs to the second contributor.

Finally it is important to note that, in order to carry on a DIP-STR analysis on the mixture, the only information needed from the main contributor's DNA is its DIP-heterozygosis or homozygosis.

3.3 An object-oriented Bayesian network (OOBN) for results of DIP-STR analyses

3.3.1 Evaluation of DNA profiling results using graphical models

Given a mixed DNA stain from two contributors, of which only one can be taken as known (say, the victim), and a suspect who shares alleles with the stain profile in some appropriate way, two main hypotheses may generally be of interest if the evaluation is addressed at source level (Cook et al., 1998). One, usually that referred to as the prosecution hypothesis (H_p), asserts that the mixture originates from the victim and the suspect (if the case is such that a suspect is available for comparative examinations). A second proposition, typically put forward by the defense (H_d), states that the mixed stain comes from the victim and an unknown person.[3] In order to assess the degree to which the profiling results allow one to discriminate between the latter two propositions, scientists need to focus on the likelihood ratio, defined as follows:

$$LR = \frac{P(E \mid H_p, I)}{P(E \mid H_d, I)}, \tag{3.1}$$

where E represents the profiling results (e.g., the genotypes of the stain, of the victim and of the suspect) and I represents the background information (i.e., the circumstances of the case). Two casework examples are proposed later in the paper, involving two different sets or profiling results. The first case covers the genotypes of a stain, a victim, and a suspect. The second case involves the genotypes of a stain and of a victim, while the suspect is supposed to be unavailable. Only his brother is available for profiling analyses.

A common interpretation of the likelihood ratio is to say that a value greater than 1 supports the prosecution hypothesis H_p, and that a value lower than 1 is in favour of the alternative hypothesis H_d. A value of 1 does not allow one to discriminate between the competing propositions of interest. As part of Bayes' theorem, the likelihood ratio connects prior odds to posterior odds in the following way:

$$\underbrace{\frac{P(H_p|E,I)}{P(H_d|E,I)}}_{\textit{Posterior odds}} = \underbrace{\frac{P(H_p \mid I)}{P(H_d \mid I)}}_{\textit{Prior odds}} \underbrace{\frac{P(E|H_p,I)}{P(E|H_d,I)}}_{\textit{Likelihood ratio}}. \tag{3.2}$$

This formula clarifies that the likelihood ratio, which is a measure of the probative value of the findings E with respect to two alternative hypotheses, is to be distinguished from the conditional degree of belief on the same hypotheses (represented by posterior odds). Notice that Equation (3.2) is a general formulation of Bayes' theorem.

When E refers to results of DNA profiling analyses, the calculation of the components of the likelihood ratio ($P(E|H_p, I)$ and $P(E|H_d, I)$) can be challenging. In relatedness testing cases, for example, the complexity of likelihood ratio formulae may be considerable, depending on parameters such as the supposed degree(s) of relatedness and the number of individuals that need to be accounted for. Moreover, formulae may vary according to genotypic configurations

[3] Here, that unknown person will be considered as unrelated to the victim.

of the target individuals. However, this computational burden can − as shown by the foundational works by Dawid et al. (Dawid et al., 2002) − be approached and safely handled through Bayesian networks to obtain the same results as those obtained by Essen-Möller's formulaic approach (focusing on posterior probabilities). In fact, Bayesian networks allow one to obtain any component defined by Equation (3.2). Thus, they prove to be a highly versatile framework that can accommodate analysts and reasoners with differing inferential interests (e.g., Kjærulff and Madsen, 2008; Taroni et al., 2006). Detailed accounts on Bayesian networks can readily be found in specialized literature (e.g., Neapolitan, 1990; Jordan, 1998; Pearl, 1988). In forensic science, they are now part of well established literature as illustrated by several reports on their application for evaluating results of forensic DNA profiling analyses (e.g., Biedermann and Taroni, 2012; Dawid et al., 2002; Cowell et al., 2006a, 2011).

For these reasons, Bayesian networks and their object-oriented extension (i.e., OOBNs) are retained as the general modeling framework in this paper. On a practical account, the models described throughout this study have been constructed with Hugin 7.4[4] (i.e., for building OOBNs and performing calculations). The forthcoming parts of this section describe the definition of two OOBNs (i.e., the main classes) to be used to approach the probabilistic evaluation of results of the particular kind of DNA profiling analyses presented earlier in Section 3.2 (i.e., the findings for the two casework examples). These two OOBNs, called here `Marker` and `Marker for brother`, will focus on, respectively, two-person mixtures with a suspect being available in the first case, and the suspect being missing in the second case. Elements of the general logic underlying the structure of the proposed OOBNs are inspired by Dawid et al. (2007), but with some definitional differences to reflect the particular mechanism of functioning of the DIP-STR typing technique.

3.3.2 The main class `Marker`

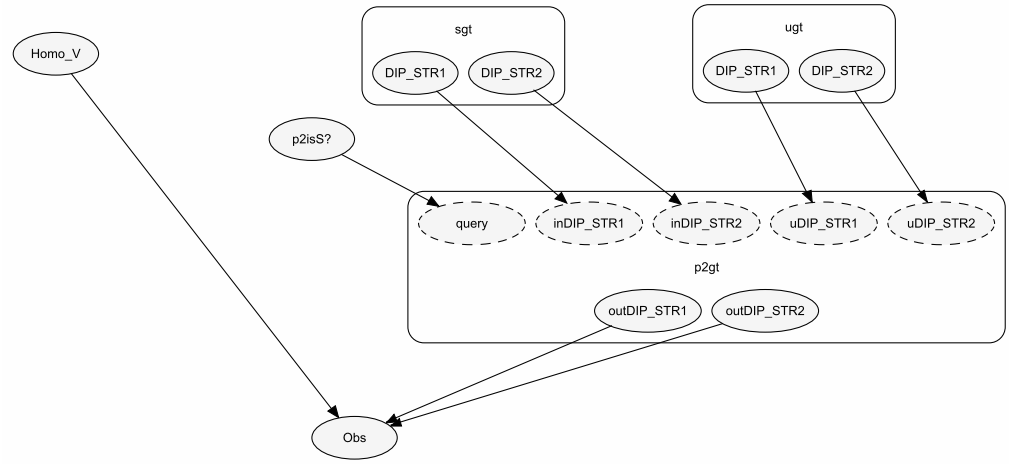


Figure 3.1: Expanded representation of the class `Marker`.

The class `Marker` (see Figure 3.1) represents the main class of the OOBN proposed to model a situation in which the evidence is given by the genotypes of the stain, of the victim and of the

[4]http://www.hugin.com.

suspect. Its main purpose is to model observations on a DNA mixture from two contributors when one of them, typically the victim, contributed more than 90% to the mixture. The remaining part is due to either the suspect or an unknown person. A collapsed version of this OOBN is given later, in Figure 3.3.

The left part of the network refers to the victim, represented by the single node **Homo_V**. This node has three states, *HomL*, *HomS* and *Hetero*, depending on whether the main contributor is homozygous or heterozygous for the DIP allele. As noted earlier in Section 3.2, this is actually the only information needed about the main contributor's genetic constitution. For purely technical reasons, the CPT of this node is completed with equal probabilities.[5]

The right part of the network (i.e., all components other than **Homo_V** and **Obs**) models the minor contributor, that could be either the suspect or an unknown person. In particular, nodes **sgt** and **ugt** are instances of the class `Genotype` (see 3.6) and represent the genotype of, respectively, the suspect and an unknown person.

The Boolean node **p2isS?** addresses the question of whether the second contributor is the suspect or an unknown person. Again, for technical reasons and invoking the same arguments as in footnote 5, equal probabilities are assigned to the CPT of this node. Node **p2gt** is an instance of the class `Pgt` (see 3.6) and represents the genotype of the actual second contributor to the mixture.

Node **Obs**, with states *La, Lb, Lab, Sa, Sb, Sab, X, nr*,[6] represents the observed (minor contributor's) DIP-STR allele(s) in the trace. These states, except *nr*, represent the results obtained when analysing the trace using the DIP primer opposite to the DIP allele of the major contributor's genotype and when one of the situations described in the first three rows of Table 3.1 holds. In turn, the state *nr* (short for 'not revealed') represents a not observed genotype, that is a result obtained whenever the major contributor is DIP-heterozygous or both contributors are DIP-homozygous of the same type.[7]

The state of the node **Obs** depends on the state of the parent node **Homo_V** that, combined with the state of the parent nodes **outDIP_STR1** and **outDIP_STR2**, determines how many STR alleles of the second contributor will appear in the result. If **Homo_V** is in the state *Hetero*, no DIP-STR profiling results can be obtained for the mixed stain. The same is the case if the first and the second contributor are DIP-homozygous of the same type. In these cases, the node **Obs** will assume the state *nr*. If the node **Homo_V** is in a state other than *Hetero*, and the second contributor is DIP-heterozygous, then only the DIP-STR allele with the DIP allele opposite to that of the first contributor is revealed. The last case is the one in which both contributors are DIP-homozygous of different type: in this case the observed minor contributor's genotype is composed by a couple of different DIP-STR alleles if it is STR heterozygous, otherwise is composed by a single DIP-STR allele as in the previous

[5] As will be shown in Section 3.4, this node will be instantiated when evaluating components of the likelihood ratio so that the initial values in its CPT are no longer of importance. The initial probabilities in the CPT are only needed from a definitional point of view, in order to build a complete model.

[6] The meaning of the letters *a*, *b* and *x* will be explained in further detail in 3.6.

[7] The term *nr* represents a situation in which no alleles of the minor contributor are revealed, due to a particular combination of the genotypes of the two contributors (see rows 3, 4 and 5 of Table 3.1). Situations in which no alleles are revealed, due to low-template traces or other problems with the PCR process, are not taken into account in the paper.

case. Table 3.2 shows part of the CPT for the node **Obs**.[8]

3.3.3 The main class `Marker for brother`

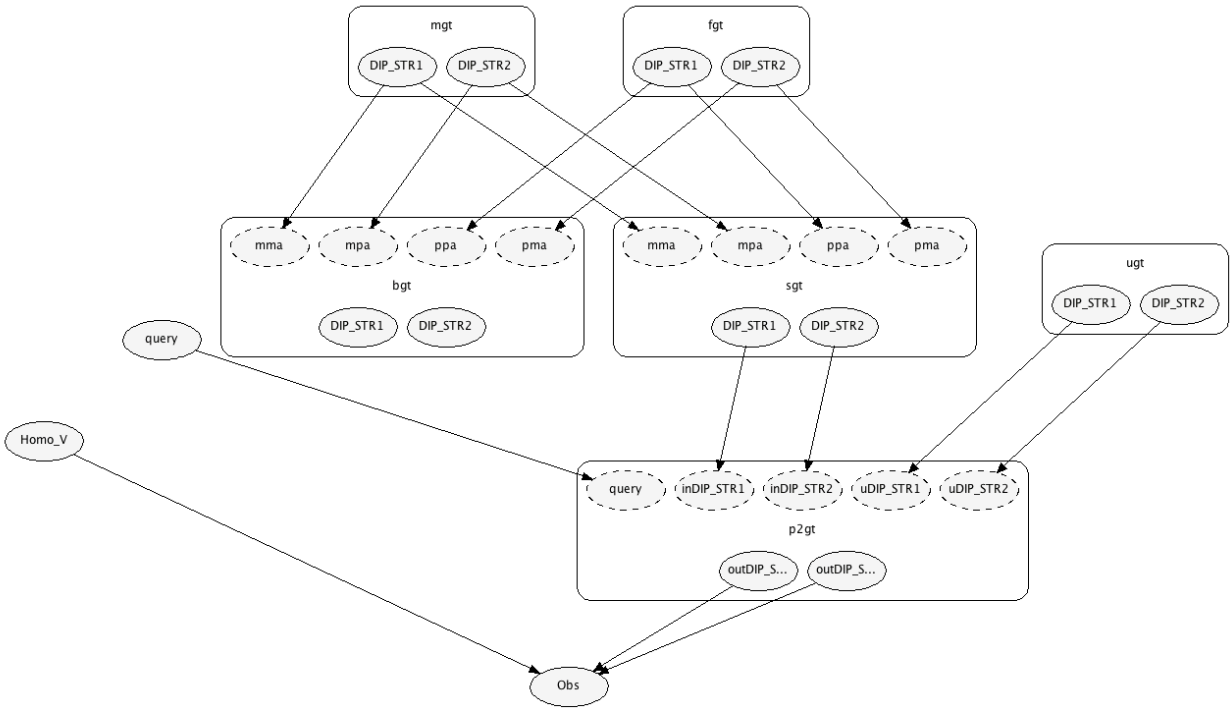


Figure 3.2: Expanded representation of the class `Marker for brother`.

The class `Marker for brother` (see Figure 3.2) represents the overall structure of the OOBN, proposed to model a situation in which profiling results consist of the genotypes of the stain, the victim, and the brother of the suspect. Its main purpose is to model profiling results for a DNA mixture from two contributors when one of them, typically the victim, contributed more than 90% to the mixture, and the suspect's DNA is not available for comparison. To deal with these missing data, nodes for the suspect's (full) brother genotype, supposed to be known, have been added. These additional nodes have been logically combined with the network through nodes representing the genotype of the brother's parents (which are also the parents of the suspect). A collapsed version of this OOBN is given later in Figure 3.5. Nodes **mgt**, **fgt** and **ugt** are instances of the class `Genotype` (see 3.6) and represent the genotype of, respectively, the suspect's mother, the suspect's father and an unknown person. Most of the nodes which are in common with the class `Marker` have the same definition, except for the node **sgt**, which, together with node **bgt**, is an instance of the class `Child` (see 3.6).

[8]The state X of the node **Obs** summarizes all cases in which the results of the analysis of the mixed stain show an allele with letter x: *Lx, Lax, Lbx, Sx, Sax, Sbx*, as explained in further detail in 3.6.

Homo_V	*HomL*																								
outDIP_STR1	*La*						*Lb*						*Lx*						*Sa*						...
outDIP_STR2	*La*	*Lb*	*Lx*	*Sa*	*Sb*	*Sx*	*La*	*Lb*	*Lx*	*Sa*	*Sb*	*Sx*	*La*	*Lb*	*Lx*	*Sa*	*Sb*	*Sx*	*La*	*Lb*	*Lx*	*Sa*	*Sb*	*Sx*	...
La	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...
Lab	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...
Lb	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...
Sa	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	1	1	1	0	0	...
Sab	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	...
Sb	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	...
X	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	...
nr	1	1	1	0	0	0	1	1	1	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	...

Homo_V	*HomS*																								
outDIP_STR1	*La*						*Lb*						*Lx*						*Sa*						...
outDIP_STR2	*La*	*Lb*	*Lx*	*Sa*	*Sb*	*Sx*	*La*	*Lb*	*Lx*	*Sa*	*Sb*	*Sx*	*La*	*Lb*	*Lx*	*Sa*	*Sb*	*Sx*	*La*	*Lb*	*Lx*	*Sa*	*Sb*	*Sx*	...
La	1	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	...
Lab	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...
Lb	0	0	0	0	0	0	0	1	0	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	...
Sa	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...
Sab	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...
Sb	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...
X	0	0	1	0	0	0	0	0	1	0	0	0	1	1	1	1	1	1	0	0	1	0	0	0	...
nr	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	...

Homo_V	*Hetero*																								
outDIP_STR1	*La*						*Lb*						*Lx*						*Sa*						...
outDIP_STR2	*La*	*Lb*	*Lx*	*Sa*	*Sb*	*Sx*	*La*	*Lb*	*Lx*	*Sa*	*Sb*	*Sx*	*La*	*Lb*	*Lx*	*Sa*	*Sb*	*Sx*	*La*	*Lb*	*Lx*	*Sa*	*Sb*	*Sx*	...
La	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Lab	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Lb	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Sa	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Sab	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Sb	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
nr	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	

Table 3.2: Partial representation of the way in which the CPT of node **Obs** is completed, depending on the possible state configurations of the parental nodes **Homo_V**, **outDIP_STR1** and **outDIP_STR2**.

3.4 Casework examples

3.4.1 General case description and DIP-STR analyses

Suppose a case in which the body of a dead women is found (Castella et al., 2013). Circumstantial evidence leads to three suspects: a man and his two sons. Other information supports the possibility of a single perpetrator, and this information is used as an assumption in the subsequent evaluation of analytical results. A relevant blood stain − denoted A here − was collected on the victim's body. Blood of the victim and of the three suspects was also available for analysis. Using the standard protocols (autosomal STR multiplex and Y-STR), the analyses led (i) to a complete autosomal STR profile matching the victim's DNA profile (without any indication of a mixed profile), and (ii) to a complete Y-STR profile, matching all the three suspects. None of the three suspects can thus be excluded as a potential contributor to the detected DNA stain. Further, it is assumed here that there are only two contributors to the DNA trace.

With the aim of discriminating between the three male suspects, it has been decided to analyze three DIP-STR loci: MID1950-D20S473, MID1107-D5S1980, MID1013-D5S490, called here Marker 1, Marker 2 and Marker 3, respectively. Table 3.3 summarizes the DIP-STR profiles of the victim and of the three suspects.

	Marker 1	Marker 2	Marker 3
Victim	S12-S13	L13-L14	S14-S14
Father	S11-S11	L13-L13	S14-S14
Son 1	S11-L12	L13-L13	S14-S14
Son 2	S11-S12	S19-L13	S14-S14

Table 3.3: DIP-STR genotypes of the victim and the three suspects.

Since the victim is DIP-homozygous (S-S, L-L and S-S) in the three selected loci, it is possible to genotype the mixture with the opposite DIP-alleles: L for Marker 1, S for Marker 2 and L for Marker 3. The results are as follows:

$$\text{Stain A:} \quad \text{Marker 1}=\{\text{L12}\}, \quad \text{Marker 2}=\{nr\}, \quad \text{Marker 3}=\{nr\}.$$

Comparing these results for DIP-STR markers to the DIP-STR genotypes of the three suspects, it can be seen that, at Marker 3, the DIP-genotypes of all suspects are compatible with the result nr for the bloodstain. At Marker 2, this happens only for the DIP-alleles of Father and Son 1. Indeed, if Son 2 contributed to the mixture, then S19 would appear in the results. Using a similar argument for Marker 1, only the DIP-genotype of Son 1 is compatible with the result for the blood stain.

Based on these DIP-STR results, Son 2 and Father can thus be excluded as contributors to the mixed stain recovered on the victim's body. This leaves only Son 1 as a potential contributor (among the individuals for which analyses have been performed), and this leads to questions of the following kind: What is the meaning of such a non-exclusion? What is the degree of support for the proposition according to which Son 1 contributed to the crime

stain? The forthcoming sections will approach such questions through the use of OOBNs and two distinct case settings. Propositions of interest can now be expressed as 'the mixed stain is made up of the DNA of the victim and Son 1' (H_p) and 'the mixed stain is made up of the DNA of the victim and an unknown person, unrelated to Son 1' (H_d).

3.4.2 Case 1: suspect available

Preliminaries

In order to build a network for illustrating the application of the OOBN modeling procedure to Case 1, instances of the class networks defined in 3.6, 3.6 and 3.6 have been combined to form the overall network called `Marker` (as introduced earlier in Section 3.3.2). Figure 3.3 shows this network in a collapsed version. In this first example, the genotype of the suspect

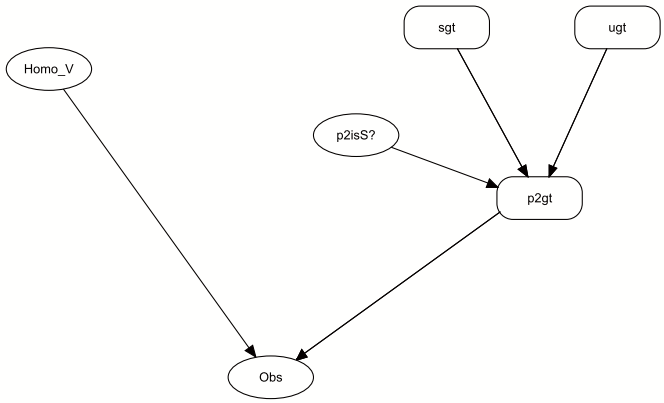


Figure 3.3: Collapsed representation of the proposed OOBN for the evaluation of DIP-STR profiling results.

is considered as known. The available items of evidence for which the probative value is to be calculated, are shown in Table 3.4.

	Marker 1	Marker 2	Marker 3
Victim	S12-S13	L13-L14	S14-S14
Suspect (Son 1)	S11-L12	L13-L13	S14-S14
Stain	L12	nr	nr

Table 3.4: Profiling results for Case 1.

Before using this OOBN for inference, one needs to decide about the definition of the states of the node **DIP_STR**, depending on the genotype of the suspect and on the results obtained after the analysis of the trace. To do this, one should take into account both the alleles observed in the suspect's profile and those in the profile of the mixture. There are 4 possible situations, summarized in Table 3.5:

- If only one allele is observed (say L12 for Marker 1) in both analyses (i.e. for the stain and the suspect), then one can refer to this with the state *La* and put probability 0

to the states relative to the alleles *Sa Sb* and *Lb* in the node **DIP_STR** of the class `DIP_STR`. This is shown in the first row of Table 3.5.

- If two alleles are observed, such as L12 L11, L12 S12 or L12 S11, the states *La Lb*, *La Sa* or *Sa Sb* can be used to represent them. In such a case, a probability equal to zero would be set for the remaining *a* and *b* alleles.

The assignment of probabilities to states *Sx* and *Lx* follows the explanations given in Section 3.6. Reminding how DIP-STR analyses work, if three (or more) different alleles are observed following the analysis of the suspect and the mixture, then the suspect can be excluded from being a contributor (under the assumption of a two person mixture). In the case of Marker 1, for example, one can use the probability of the allele L12 for the state *La* of **DIP_STR**, and the probability of the allele S11 for the state *Sb*. The probability for states *Lb* and *Sa* will be set to 0, while the probability for the states *Lx* and *Sx* will be calculated by adding the probabilities of all the other L- (and S-) alleles. A summary of this is given in row three of Table 3.5.

The described procedure of assigning letters *a*, *b* and *x* and the correct probabilities to the states of node **DIP_STR** may be viewed as time-demanding or prone to errors. It is for this reason that an R function, written with the package RHugin Konis (2010) is available for the interested readers (requests by e-mail to the Corresponding author). The function is called `Dipstr_LR` and is of the following form:

```
Dipstr_LR <- function(Marker_name, vgt1, vgt2, sgt1, sgt2, Obs1, Obs2)
```

where one has only to specify the marker to be considered, the genotype of the victim, of the suspect and the two alleles observed in the trace.

Results from the stain	**Suspect genotype**	**La**	**Lb**	**Lx**	**Sa**	**Sb**	**Sx**
L12	L12 L12	γ_{L12}	0	$\sum_{j\in J\setminus\{12\}}\gamma_{Lj}$	0	0	$\sum_{k\in K}\gamma_{Sk}$
L12	L12 S12	γ_{L12}	0	$\sum_{j\in J\setminus\{12\}}\gamma_{Lj}$	γ_{S12}	0	$\sum_{k\in K\setminus\{12\}}\gamma_{Sk}$
L12	L12 S11	γ_{L12}	0	$\sum_{j\in J\setminus\{12\}}\gamma_{Lj}$	0	γ_{S11}	$\sum_{k\in K\setminus\{11\}}\gamma_{Sk}$
L12 L11	L12 L11	γ_{L12}	γ_{L11}	$\sum_{j\in J\setminus\{11,12\}}\gamma_{Lj}$	0	0	$\sum_{k\in K}\gamma_{Sk}$

Table 3.5: Possible probability assignments for the states La, Lb, Lx, Sa, Sb et Sx of the node **DIP_STR**. J is the set of all the possible STR alleles linked to the DIP allele L, that can be present in the marker of interest and K is the set of all the possible STR alleles linked to the DIP allele S. The choice of the numbers 11 and 12 serves the sole purpose of illustration. An analogous table can be built to model situations in which DIP alleles of the type S are observed in the stain.

Likelihood ratios for DIP-STR profiling results: instantiating the OOBN

To obtain a likelihood ratio, scientists need to evaluate two conditional probabilities: $P(E|H_p, I)$ and $P(E|H_d, I)$. Here, the variable E refers to the results for the trace (E_{obs}) and the geno-

type of the victim and the suspect. The latter two genotypes will be denoted E_g for short. The likelihood ratio (LR) can thus be written as follows:

$$LR = \frac{P(E \mid H_p, I)}{P(E \mid H_d, I)} = \frac{P(E_{obs}, E_g \mid H_p, I)}{P(E_{obs}, E_g \mid H_d, I)} = \frac{P(E_{obs} \mid H_p, E_g, I)P(E_g|H_p, I)}{P(E_{obs} \mid H_d, E_g, I)P(E_g|H_d, I)} \tag{3.3}$$

$$= \frac{P(E_{obs} \mid H_p, E_g, I)}{P(E_{obs} \mid H_d, E_g, I)}. \tag{3.4}$$

The last equality is obtained by invoking the assumption that the victim's and the suspect's genotype (represented by E_g) do not depend on whether the suspect is or is not a contributor to the mixed stain, given the background information I.

To obtain a value for the numerator of the likelihood ratio, instantiations should be made in the nodes **Homo_V?**, **DIP_STR1** and **DIP_STR2** of the class `sgt`, and **p2isS?**. In particular, the latter node needs to be set to the state *True*, because under H_p it is the suspect who is assumed to be the second contributor. The probability of observing a given DIP-STR configuration for the mixed stain − under this conditioning − is then read from the node **Obs**. In turn, a value for the denominator is obtained by instantiating the node **p2isS?** to the state *False* and then, again, reading the required conditional probability in the node **Obs**.[9] The ratio between the two numbers thus found gives the likelihood ratio.

Figure 3.4 illustrates these instantiations and propagations in terms of the OOBN `Marker`, used to represent the results obtained in the currently discussed casework example for Marker 1. Figure 3.4(a) illustrates how to obtain the numerator, whereas Figure 3.4(b) illustrates the evaluation of the denominator. Notice that the instantiations made in the nodes **DIP_STR1** and **DIP_STR2** of the class `sgt` are not visually displayed because they are operated inside instance nodes. Prior to using the OOBN for these propagations, one needs to tailor the probability table of the node **DIP_STR**. In view of the observation that, for Marker 1, the mixed stain shows L12 and the suspect has the genotype S11-L12, probabilities as shown in row three of Table 3.5 need to be specified. On the basis of internal data collected at the authors' institution (on 100 individuals), the following vector of probabilities was assigned:

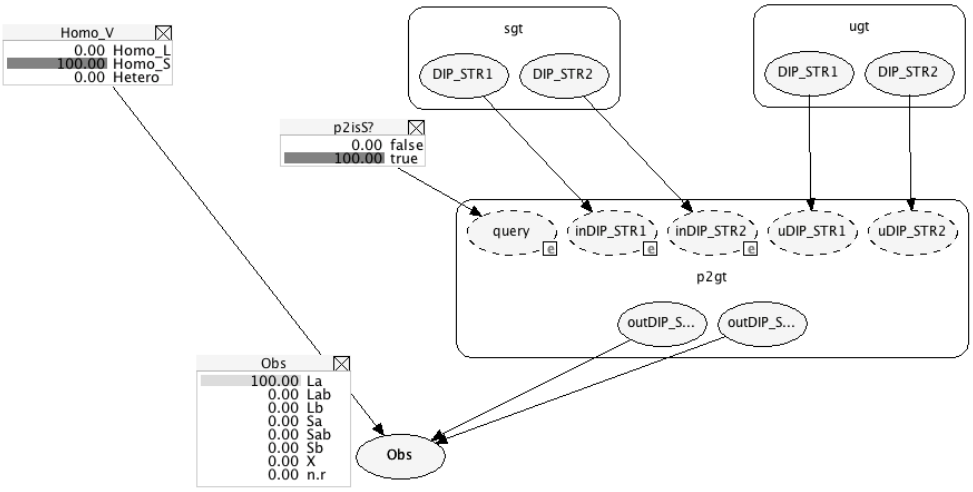
$$P(\{La, Lb, Lx, Sa, Sb, Sx\}) = \{0.181, 0, 0.208, 0, 0.259, 0.352\}.$$

The probabilities for the state *La* of the node **Obs** found in the described way lead to the following likelihood ratio:

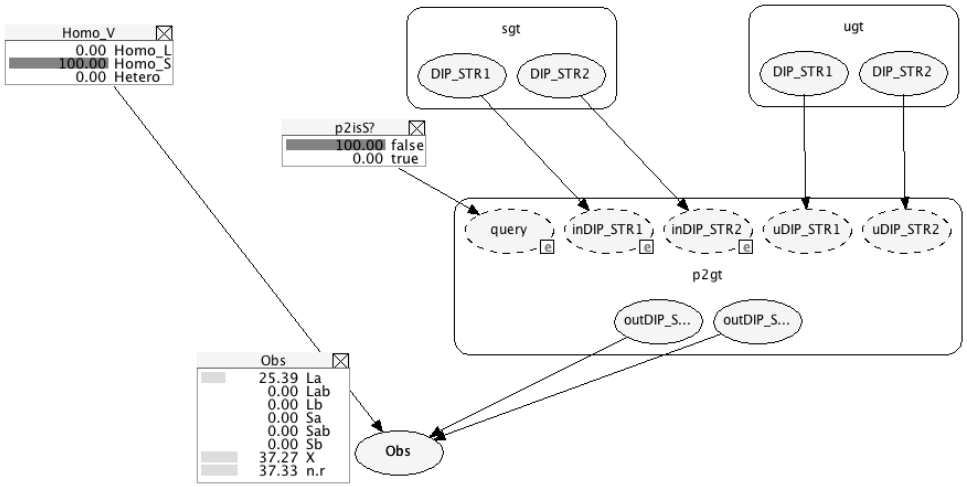
$$LR = \frac{P(E_{obs} \mid H_p, E_g, I)}{P(E_{obs} \mid H_d, E_g, I)} \tag{3.5}$$

$$= \frac{P(\mathbf{Obs} = La \mid \mathbf{P2isS?} = True, \mathbf{Homo\_V} = HomS, \mathtt{sgt} = \{La, Sb\}, I)}{P(\mathbf{Obs} = La \mid \mathbf{P2isS?} = False, \mathbf{Homo\_V} = HomS, I)} \sim \frac{1}{0.2539} \sim 3.95. \tag{3.6}$$

[9] Notice that given **p2isS?**=*False*, information about the genotype of the suspect becomes irrelevant. That is, any instantiation made in the nodes modeling the suspect's genotype will not affect the probabilities obtained at the node **Obs**.



(a) Under the proposition H_p, the node **p2isS?** is set to *True*.



(b) Under the proposition H_d, the node **p2isS?** is set to *False*.

Figure 3.4: OOBN `Marker` used to evaluate profiling results obtained for Marker 1 where (a) represents the view under the first proposition (i.e., H_p) and (b) represents the the view under the alternative proposition (i.e., H_d).

The numerator of this result can readily be understood. If the suspect, whose genotype is L12-S11, is truly the second contributor, and the analyst analyses the trace using primers for the L-DIP, then it can reasonably be expected that L12 will be detected in the stain.[10] Assuming no disturbing or otherwise complicating factors during the analyses, a numerator of 1 can thus be found. For the denominator, further considerations are required. Under the assumption of a contributor other than the suspect, one needs to consider several possible genotypes. In fact, the second contributor could have any of the following genotypes: $La-La, La-Sb, La-Sx$, for $a = 12$, $b = 11$ and $x \neq 11, 12$. The value of the denominator is thus given by the sum of the probabilities of these genotypes. This leads to the following expression of the likelihood ratio:

$$LR = \frac{P(E_{obs} \mid H_p, E_g, I)}{P(E_{obs} \mid H_d, E_g, I)} = \frac{1}{\gamma_{La}^2 + 2\gamma_{La}\gamma_{Sb} + 2\gamma_{La}\gamma_{Sx}} \sim \frac{1}{0.2539} \sim 3.95. \tag{3.7}$$

This result demonstrates that the OOBN-output is not arbitrary, but can be reproduced as a result of logical considerations. It is also worth mentioning that the result can also be related to existing literature on qualitative mixture assessment as described by Weir et al. (Weir et al., 1997) (based on Evett (Evett et al., 1991)). Although the formulae derived in these references are intended to evaluate traditional STR profiling results, their underlying logic also applies to DIP-STR results: the aim is to find the probability that a given number of persons possess − in combination − particular alleles (here: DIP-STR alleles).

In the same way as outlined here above, one can also find likelihood ratios for DIP-STR profiling results on Marker 2 and Marker 3. Table 3.6 summarizes these results, as well as the overall likelihood ratio. The latter value is obtained by multiplication because the DIP-STR markers are assumed to be independent, in the same way as traditional STR markers. For the time being, the results summarized in Table 3.6 should be taken as provisional because the collection of relevant data, in a more extensive form, is still underway.

Marker:	Marker 1	Marker 2	Marker 3	Combined
Likelihood ratio:	3.9	2.2	1.8	~ 15

Table 3.6: Summary of the likelihood ratios obtained for profiling results on three DIP-STR markers for Case 1, rounded to one decimal.

The probabilistic analyses conducted in this section, such as Equation 3.7, may appear elementary. However, they may become tedious if they need to be done manually. The reason for this is that, for each marker, distinct allelic configurations may be observed so that the formulaic development may take different forms. The advantage of using an OOBN thus becomes immediately clear. Except for the input values (i.e., the initial numerical specification), the model structure remains constant. Moreover, the analyst can confine computations entirely to the model. Thus, the use of an OOBN could also help to make evaluative procedures less prone to possible errors. This is further clarified in the next casework example (Section 3.4.3), where the suspect's genotype is supposed to be unavailable. Generally, formulae readily become more complicated in such settings, depending on the degree of relatedness between the suspect and the typed individuals.

[10] Recall that no alleles from the victim will be detected because he has only S-DIP alleles.

3.4.3 Case 2: missing suspect

This example uses some of the data of Case 1. As a main difference, it is supposed that the genotype of the suspect, as well as that of his father, are not available. Table 3.7 provides a summary of the available profiling results.

	Marker 1	Marker 2	Marker 3
Victim	S12-S13	L13-L14	S14-S14
Brother (Son 2)	S11-S12	S19-L13	S14-S14
Stain	L12	nr	nr

Table 3.7: Profiling results for Case 2.

The network specifications and the definition of the states of the different nodes of the class `Marker for brother` are as described in Section 3.3.3, and displayed in collapsed form in Figure 3.5. Table 3.8 summarizes the likelihood ratios obtained for Case 2.

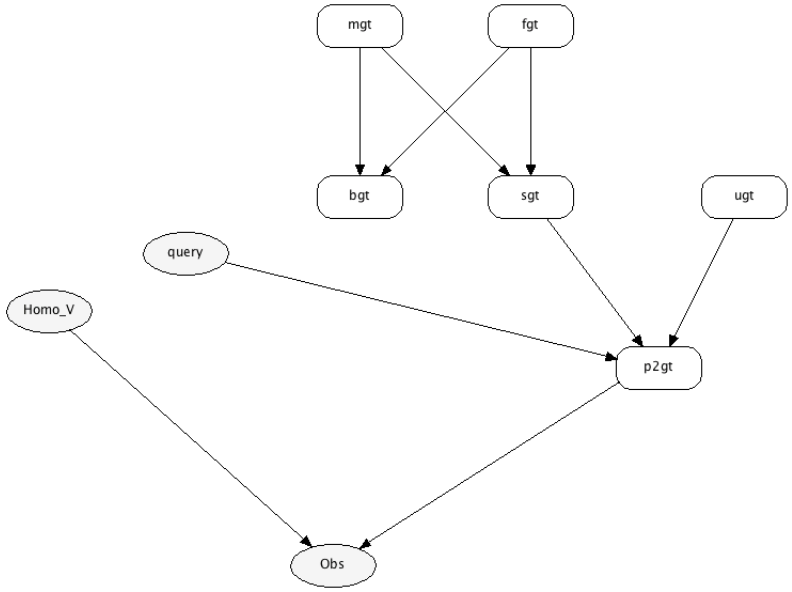


Figure 3.5: Collapsed representation of the class `Marker for brother`.

Marker:	Marker 1	Marker 2	Marker 3	Combined
Likelihood ratio:	0.6	0.6	1.4	~ 0.5

Table 3.8: Summary of the likelihood ratios obtained for profiling results on three DIP-STR markers for Case 2, rounded to one decimal.

3.4.4 A note on the likelihood ratio results

In Case 1, an overall likelihood ratio of about 15 was obtained. This means that the findings (i.e., results for three markers) support the proposition according to which the victim and

the suspect Son 1 are the two contributors to the mixture. Although this likelihood ratio is rather moderate, it is useful to note some further aspects of this result. First, when the case was actually examined, only three DIP-STR markers had been used. At the moment, a panel of 9 markers is available which could allow one to obtain higher likelihood ratios. Moreover, the likelihood ratio result is higher than the likelihood ratio of one that would be obtained in situations where the classical STR method does not yield any profiling output, which typically occurs with strongly imbalanced mixtures. Finally, one should also consider that even in presence of a *nr* result for the stain, when the victim is DIP homozygous a likelihood ratio higher than one is obtained. This is so because the result indicates that the victim and the second contributor are DIP homozygous of the same kind (both S-S or L-L).

Often, there are unfortunate expectations in the field that forensic DNA must necessarily be (highly) probative, but this should not distract us from devoting attention to alternative profiling techniques that usefully complement the broad range of approaches available to the forensic practitioner. Moreover, the resulting likelihood ratio can be aggregated with the result obtained from Y-STR profiling analyses, using the same couple of target propositions.

In Case 2, an overall likelihood ratio of about 0.5 was obtained. This means that the findings (i.e., results for three markers) slightly support the proposition according to which Son 1 is not a contributor to the mixture. This result is not unreasonable since markers in which the brother's genotype is incompatible with the stain results tend to lead to a LR$<$ 1: since the suspect is genetically close to his brother, this observation will also hold for the suspect.

3.5 Discussion and conclusions

Unbalanced DNA mixtures are problematic for traditional STR profiling analyses, in particular when the proportion between the DNA of the two contributors is more extreme than 1:10 (Clayton and Buckleton, 2005). Cases of sexual assaults (where the victim's DNA is predominant and that of the aggressor is present only as a minor quantity) or cases of micro chimerism during pregnancy (where minute quantities of fetal DNA are present in maternal blood) are typical examples for situations in which stains of this type may be found. To cope with this constraint, recent developments focused on alternative analytical methods using a new compound marker, formed by a STR marker coupled to a DIP (Castella et al., 2013). A particular feature of DIP-STR markers is that, whenever they can be analyzed, they can detect alleles directly related to the second contributor. However, the successful detection of DIP-STR alleles in an unbalanced mixed stain depends on how the DIP-STR genotypes of the stain contributors compare to each other, as there may be situations in which none or only part of the target DNA of the individual of interest (i.e., different from the assumed known contributor, such as a victim) can be detected.

In order to make this novel DNA profiling technique applicable in forensic contexts, one needs to be able to assess the meaning of particular profiling results with respect to selected competing propositions. Examples include 'the victim and the suspect contributed to this DNA mixture' versus 'the victim and an unknown person contributed to this DNA mixture'. This paper has investigated the use of graphical probability models (i.e., Bayesian networks),

in particular OOBNs, to address questions of this kind. OOBNs have been chosen because they allow one to derive a concise representation of the genotypic configuration of the various (assumed) contributors as well as the mixed stain. Most importantly, such graphical networks allow one to depict the way in which the assumed contributors' genotypes relate to that of the crime stain. On a computational account, such models also allow their user − and this is one of the main features of OOBNs − to find the components of likelihood ratios that express the probative value for particular findings (i.e., DIP-STR profiling results). OOBNs can thus help scientists to deal with the (often complex) calculations that are encountered with DNA mixtures. For example, an OOBN will require only minimal initial specifications in order to approach the typing results for a given marker. Typically, these initial specifications will relate to the probabilities assigned for the various alleles, but with regards to the qualitative graph structure, the model should not require any changes. This is different for purely formulaic approaches to evaluation because these may take various different forms, depending on the particular profiling results (for both the potential contributors as well as the mixed crime stain), and may thus be less practical in their application − eventually also more prone to error (if they need to be done manually). Moreover, as pointed out in Dawid et al. (2002), the advantage of using a graphical probabilistic approach becomes evident in cases where genetic information of further individuals (other than the suspect) need to be considered (typically when the suspect is missing, and information on his genotype is not available). A purely arithmetic solution to such problems may become increasingly challenging. Such a situation is encountered in Case 2. The corresponding OOBN shows how the brother's genotype can be considered through a very straightforward modification of the OOBN's structure.

The rather moderately sized likelihood ratio values obtained for the reported casework examples should not be interpreted as a limiting factor in principle. Indeed, it is worthwhile to emphasize that (i) the described profiling technique (DIP-STR) works with particularly high reliability under special circumstances implied by unbalanced mixtures (at least in the case in which the two contributors are of the same gender, or the minor contributor is a female), (ii) potential stain contributors could be excluded, and (iii) the probative value for non-excluded individuals can be characterized probabilistically. Future research in the authors' institution will focus on investigating further markers of this kind, as well as the generation of relevant population data to improve the numerical specification of the proposed OOBN-approach.

3.6 Acknowledgements

This research was supported by the Swiss National Science Foundation, through grant no. 105311- 1445570.

Appendix A. Details on the OOBN to model DIP STR results

This appendix describes in details the classes which are contained in the main class `Marker` (see Section 3.3.2)

The class `DIP_STR`

Figure 3.6: Representation of the class `DIP_STR`.

The class `DIP_STR` (see Figure 3.6) consists of a single output node **DIP_STR** whose states represent the different DIP-STR alleles, denoted here *La, Lb, Lx, Sa, Sb* and *Sx*. The CPT contains the probability of occurrence of these alleles in the relevant population. In order to be coherent with a Bayesian approach, a Bayesian estimation is used here for the allele proportions, based on a prior Dirichlet distribution for those proportions (e.g., Evett and Weir, 1998; Taroni et al., 2010; Brandwein and Strawderman, 2005).[11] At this point, the assumption of non-independence between DIP and STR markers, made earlier in Section 3.2, becomes explicit. This understanding is conveyed by using only a single node **DIP_STR**, rather than two separate nodes. Letters *a*, *b* and *x* are used instead of the list of the actual STR allele numbers in order to use the model for different markers, and to facilitate computational tasks.

Note that, according to currently available data, there are loci with more than six different alleles. To handle this, states *La*, *Lb*, *Sa* and *Sb* are used to represent the two alleles that, at most, could be observed with the DIP-STR method,[12] while states *Lx* and *Sx* represent all the alleles that, in principle, may appear in that locus but are not actually observed. This means that, before deciding which alleles to associate with the states of the node **DIP_STR**, one will consider the results of the analysis performed on the DNA mixture (see Section 3.4.2). The probability of states *Lx* (and *Sx*) is set as the sum of all the probabilities of the unobserved L-STR alleles (S-STR alleles). As already explained, the state X of the node **Obs** is reserved for results that show an allele corresponding to letter *x*. Practically, this will not be the case, since *x* represents the not observed STR alleles, so that the state X may be said to have a dummy function. It is needed only for structural reasons, in order to complete the network from a definitional point of view.

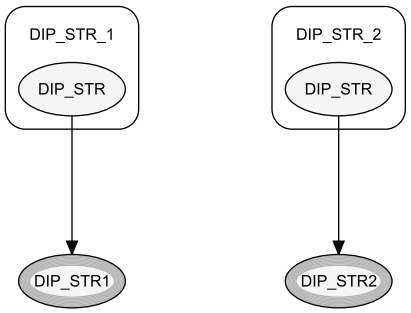


Figure 3.7: Representation of the class `Genotype`.

The class `Genotype`

The class `Genotype` (see Figure 3.7) is used in the main class `Marker` to represent genotypes of different individuals involved in the case, when there is no need to include an explicit representations of their parents. This class is itself composed by instance nodes, namely **DIP_STR1** and **DIP_STR2**. These are instances of the class `DIP_STR` and represent the allelic constitution, on the two chromosomes, of the given person. The class `Genotype` also contains output nodes, called **DIP_STR1** and **DIP_STR2**, that are copies of their parent nodes.[13] This class is used in the main class `Marker` through two instances (**sgt** and **ugt**) representing, respectively, the genotype of the suspect and of an unknown person. It is also used in the main class `Marker for brother` through the instances **mgt**, **fgt**, and **ugt**.

The class `Child`

Figure 3.8: Representation of the class `Child`.

The class `Child` (Figure3.8) is used in the main class `Marker for brother` to represent genotypes of individuals for which it is necessary to include their parents explicitly (i.e., the genotype of a person of interest is presented as a child variable depending on the genotypic configuration of the parents). This class contains (i) two output nodes, called **DIP_STR1** and **DIP_STR2**, that represent the allelic constitution (on the two chromosomes) of the given person, (ii) two input nodes **mpa** and **mma** which represent the two alleles possessed by the mother (one of which is inherited by the child), and (iii) two input nodes **fpa** and

[11] In what follows, the Bayesian estimate for the Li allele proportion is referred to as γ_{Li}.

[12] Four different states are needed even if at most two alleles can be observed at a given marker. This is because − taking into account the two different DIP alleles − four different combinations can appear.

[13] The purpose of this is to have the nodes **DIP_STR1** and **DIP_STR2** as output nodes in the main class `Marker`.

fma which represent the two alleles possessed by the father (one of which is inherited by the child). This class is used in the main class `Marker for brother` through two instances (**sgt** and **bgt**) representing, respectively, the genotype of the suspect and of the brother of the suspect.

The class `Pgt`

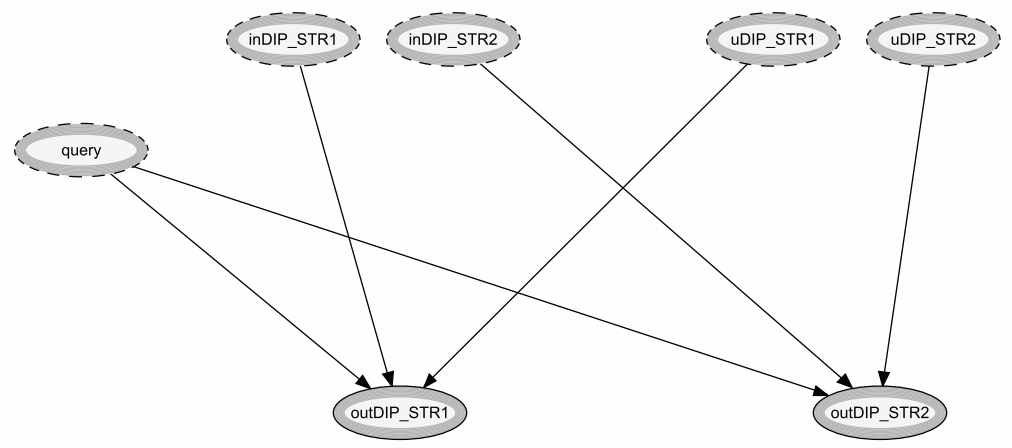


Figure 3.9: Representation of the class `Pgt`.

The class `Pgt` (see Figure3.9) is used to represent the allelic configuration, on the two chromosomes, of the actual second contributor to the mixture. It is composed by input and output nodes. The input nodes are **query**, that is a Boolean node, bounded to the node **p2isS?** of the external network (e.g., Figure 3.1). The nodes **inDIP_STR1** and **inDIP_STR2** are bound, respectively, to the nodes **DIP_STR1** and **DIP_STR2** of the instance **sgt** of the class `Genotype`. The nodes **uDIP_STR1** and **uDIP_STR2** are related, respectively, to the nodes **DIP_STR1** and **DIP_STR2** of the instance **ugt** of the class `Genotype`. The output nodes **outDIP_STR1** and **outDIP_STR2** are copies of the nodes **inDIP_STR1** and **inDIP_STR2** if the node **query** is in the state *True*, otherwise they are copies of the nodes **uDIP_STR1** and **uDIP_STR2**. This represents the understanding that, if the second contributor is the suspect (i.e., the node **p2isS?** is in the state *True*), then the genotype of the second contributor (modeled by **outDIP_STR1** and **outDIP_STR2**) should reflect the genotype of the suspect. Otherwise it should be equal to the genotype of an unknown person from the relevant population (represented by the nodes **uDIP_STR1** and **uDIP_STR2**). The CPT of the nodes **outDIP_STR1** and **outDIP_STR2** thus are completed as follows:

For $i = \{1, 2\}, j = \{La, Lb, Lx, Sa, Sb, Sx\}$, it holds

- $P(\mathbf{outDIP\_STR}i = j \mid \mathbf{uDIP\_STR}i = j, \mathbf{query} = False) = 1$
- $P(\mathbf{outDIP\_STR}i = j \mid \mathbf{inDIP\_STR}i = j, \mathbf{query} = True) = 1$

Chapter 4

An investigation of the potential of DIP-STR markers for DNA mixture analyses

This chapter is based on:
Cereda, G., Biedermann, A., Hall, D., and Taroni, F. (2014). An investigation of the potential of DIP-STR markers for DNA mixture analzses. *Forensic Science International: Genetics*, 11:229 - 240.

Abstract

The genetic characterization of unbalanced mixed stains remains an important area where improvement is imperative. In fact, using the standard tools of forensic DNA profiling (i.e., STR markers), the profile of the minor contributor in mixed DNA stains cannot be successfully detected if its quantitative share of DNA is less than 10% of the mixed trace. This is due to the fact that the major contributor's profile "masks" that of the minor contributor. Besides known remedies to this problem, such as Y-STR analysis, a new compound genetic marker that consists of a Deletion/Insertion Polymorphism (DIP) linked to a Short Tandem Repeat (STR) polymorphism, has recently been developed and proposed Castella et al. (2013). These novel markers are called DIP-STR markers. This paper compares, from a statistical and forensic perspective, the potential usefulness of these novel DIP-STR markers (i) with traditional STR markers in cases of moderately unbalanced mixtures, and (ii) with Y-STR markers in cases of female-male mixtures. This is done through a comparison of the distribution of 100,000 likelihood ratio values obtained using each method on simulated mixtures. This procedure is performed assuming, in turn, the prosecution's and the defence's point of view.

4.1 Introduction

The common way to analyse DNA mixtures for forensic purposes is to use the Polymerase Chain Reaction (PCR) and STR markers (Butler, 2011). One of the limitations of this method is that it does not work successfully if the proportion of the DNA quantities of the two contributors is more extreme than 1:10 (Clayton and Buckleton, 2005). Here, the threshold of 10% is retained as the limit of detection of the minor DNA for blood:blood mixtures. This value varies depending on the types of biological fluids which constitute the mixture and the specific combination of genotypes present in the mixture (as reported in (Applied Biosystems, 2012)) and should be assessed in the validation procedure (Butler, 2011). Mixtures with such extreme proportions are referred to in this paper as 'extremely unbalanced mixtures', opposed to 'moderately unbalanced mixtures', that are mixtures for which the proportion of DNA of each contributor is less extreme than 1:10. Situations involving extremely unbalanced mixtures are quite common, such as in cases of sexual assaults when the victim's DNA is largely predominant or cases of microchimerism during pregnancy (where minute quantities of fetal DNA are present in maternal blood). To address constraints implied by these kind of mixtures, Y-STR markers are widely adopted (Roewer, 2009), with the limitation that they provide information on the minor contributor only if that individual is male and the major contributor female. To address both the constraints of mixture imbalance and contributors' gender mismatch, an alternative analytical method has recently been developed and proposed (Castella et al., 2013). It is based on the use of new compound markers, each formed by an STR marker coupled to a marker in which a Deletion/Insertion Polymorphism (DIP) (Weber et al., 2002) is known to be present. So far a panel of 9 markers has been provided, called DIP-STR markers.

An object-oriented Bayesian network for the assessment of profiling results obtained with this novel technique has been developed (Cereda et al., 2014b). This network approach allows one to calculate a likelihood ratio for mixtures of two contributors, when the major contributor's genotype is known and the two competing hypotheses are 'the minor contributor is the suspect' (H_p) and 'the minor contributor is an unknown person, unrelated to the suspect' (H_d).

This paper aims to compare, from a statistical and forensic perspective, the potential usefulness of these novel DIP-STR markers (i) with traditional STR markers in cases of moderately unbalanced mixtures, and (ii) with Y-STR markers in cases of female-male mixtures. Section 4.2 starts with a brief introduction to the characteristics of the DIP-STR method along with the specification of the chosen STR and Y-STR marker systems. Next, Section 4.3 will present the interpretative model and the probabilistic tools (among which are graphical models) used to produce (through simulation techniques) likelihood ratio (LR) results for the three methods. Section 4.4 compares the distributions of the likelihood ratio results for DIP-STR and classical STR, and for DIP-STR and Y-STR. Section 4.5 focuses on the study of potential usability of the methods, that is the percentage of cases in which they are useful for the purpose of the investigation. The last Section 4.6 presents a discussion and conclusions, while the Appendix provides additional tables and figures.

4.2 Genetic background

This section briefly introduces the reader to the genetical background of DIP-STR markers. It also specifies the chosen STR and Y-STR marker systems. Particular features of the three methods, which are relevant for the understanding of the forthcoming sections, are also mentioned.

4.2.1 DIP-STR markers

DIP-STR markers were recently proposed as novel type of genetic markers (Castella et al., 2013). The novelty consists in pairing a Deletion/Insertion Polymorphism (DIP) (Weber et al., 2002) with a standard STR, to form a superlocus where the two composing loci are not independent because they are so close on the chromosomes (less than 500bp apart) that they cannot recombine, but independence can be assumed between the different DIP-STR markers. Two alternative allele-specific primers overlapping the DIP locus are designed, denoted L-DIP primer and S-DIP primer (L for *long* or S for *short*). Each of these is to be used together with a primer downstream the STR region. They are useful for mixtures of any unbalance proportion (DIP-STR genotypes of minor contributors were successfully typed at a ratio up to 1:1000 (Castella et al., 2013)) and where one contributor can be assumed as known, but they present a particular interest for extremely unbalanced mixtures, when the use of STR primers leads to masking of the minor contributor's genotype by the major contributor's genotype. This is due to the fact that the STR primers are loci specific. Two contributors necessarily have alleles from the same locus, although of possibly different lengths (i.e., repeat numbers), but STR markers do not differentiate between different alleles of the same locus in case of extremely unbalanced mixtures. In practice it is observed that annealing occurs mainly with those alleles that are present in predominant quantity, so that DNA of a minor contributor will not be successfully replicated. Due to the allele specificity of DIP-STR markers, DIP-STR genotyping allows the selected amplification of the unknown contributor's DNA, as long as it has alleles that are absent in the known contributor's genotype. For the purposes of this article, the known contributor is considered as the major one.

A first important feature of this set of markers concerns the exhaustiveness of the information that can be retrieved about the minor contributor, which depends on the combination of DIP alleles of the two contributors. This is why an initial step in the analysis consists in genotyping the major contributor's DNA, in order to know which DIP-primer to use for each locus of the mixture: if, at a particular locus, the major contributor is homozygous for the DIP alleles (i.e., S-S or L-L), the DIP-primer corresponding to the other DIP allele (L if the major contributor is S-S, S if the major contributor is L-L) will be used. Note that in case the major contributor is heterozygous for the DIP alleles (i.e., S-L), none of the DIP-primers is worth to be used at that particular locus. The best scenario is when the DNA of the major and the minor contributor are homozygous for different DIP alleles (i.e., one S-S and the other L-L, or viceversa). In this case, the possible results can show either two different minor DNA haplotypes or one, depending on the STR-homozygosity or heterozygosity of the minor contributor. On the other hand, when the major contributor is DIP-homozygous and the minor contributor is DIP-heterozygous, only one haplotype of the

minor DNA can be retrieved (i.e., the one with the DIP allele opposite to the DIP allele of the major contributor's DNA). A limitation of this method is that, when the predominant DNA is DIP-heterozygous or both contributors are DIP-homozygous of the same type, it is not possible to obtain any result from the analysis of the mixture, since both DIP primers (S and L), if used, will anneal to the major contributor's DNA. Table 4.1 summarises the possible outcomes. However, it is important to point out that even in those situations for which no alleles of the minor contributor are obtained, if the major contributor is DIP homozygous, some information about the minor contributor are nevertheless obtained, because it indicates that the minor contributor has the same DIP-homozygosity as the major contributor (both S-S or L-L).

DIP genotype of major contributor	**DIP genotype of minor contributor**	**DIP-STR results**
S-S	S-S	No results
	L-L	Complete genotype of the minor contributor
	S-L	Only the L DIP-STR allele
L-L	S-S	Complete genotype of the minor contributor
	L-L	No results
	S-L	Only the S DIP-STR allele
S-L	S-S	No results
	L-L	
	S-L	

Table 4.1: Informativeness of the different genotypic DIP-STR configurations. This table represents a single locus configuration, and the results in the last column are obtained using the DIP primer opposite to the DIP primer of the major contributor.

A first panel of DIP-STR markers,[1] was introduced in Castella et al. (2013): they are referred to in this paper as Marker 1, Marker 2, ..., Marker 9, respectively. Data from 103 unrelated Swiss individuals are used here for a Bayesian estimation of the allelic proportions at each of these markers. For further information on this method, see also Cereda et al. (2014b).

4.2.2 STR markers

STR markers are routinely used to genotype DNA traces (Butler, 2011). For the purpose of the current discussion, it is important to note that in case of extremely unbalanced mixtures, the use of STR markers generally does not allow one to be aware of the presence of a mixture, since the minor contributor's profile is masked by that of the major contributor.

[1]MID1013-D5S490, MID1950-D20S473, MID1107-D5S1980, rs11277790-D10S530, rs60194384-D15S1514, rs67842608-D5S468, rs66679498-D2S342, rs10564579-D3S1282, rs35708668-D5S2045

The 16 STR markers considered here are those of the kit AmpFℓSTR®NGM SElectTM NGMSelect^{TM}(Green et al., 2013). Data from 200 Swiss unrelated individuals will be used for a Bayesian estimation of the allelic proportions at each of these markers.

4.2.3 Y-STR markers

The term Y-STR locus designates an STR locus situated on the Y-chromosome (Butler, 2005). Y-STR markers are often used in forensic casework (e.g. Roewer, 2009; Roewer et al., 1992), in particular for their capacity to reveal male-specific Y-STR alleles in male/female DNA mixtures, even if extremely unbalanced. This makes them very useful in case of extremely unbalanced mixtures in which the major contributor is a female and the minor contributor is a male. However, in case the major contributor is male they are not useful.

Another drawback of the use of Y-STR markers is that the interpretation of Y-STR results is complicated by their haploidy and patrilineal inheritance, because male relatives will share the same Y-STR profile, even over several generations (if no mutations occur). Practically, this means that even in presence of a correspondence between the Y-STR profile of the crime stain and that of the suspect, his patrilineal relatives cannot be excluded as donors of the stain. Recently, a panel of 13 rapidly mutating (RM) Y-STR markers has been identified (Ballantyne et al., 2012), which successfully differentiates between closely and distantly related males. However, both the classical and the RM Y-STR techniques are useful only if the major contributor is a women and the minor contributor is a man.

It is important to mention that, due to the lack of recombination, Y-STRs form a single haplotype (i.e., the different markers cannot be considered independent).

The discussion presented in this paper refers to the PowerPlex® Y System Thompson et al. (2013). Data from 150 Swiss male unrelated individuals Haas et al. (2006) are used for a Bayesian estimation of the Y-STR haplotype proportions.

4.3 Interpretative model

Given a DNA mixture of two contributors, of which only one can be taken as known (say, the victim), and a suspect (available for comparative analyses) who shares alleles with the stain profile in some appropriate way, the two propositions of interest are typically addressed at 'source level' (Cook et al., 1998) and can be expressed as follows: H_p (usually referred to as the prosecution hypothesis), which asserts that the mixture originates from the victim and the suspect, and H_d (the defence hypothesis), which states that the mixed stain comes from the victim and an unknown person unrelated to the suspect. In order to assess the degree to which the profiling results allow one to discriminate between these two propositions, scientists should focus on the likelihood ratio, defined as follows:

$$LR = \frac{P(E \mid H_p, I)}{P(E \mid H_d, I)}. \tag{4.1}$$

This is a ratio of two probabilities P, where E represents the profiling results (i.e., the genotypes of the stain, of the victim and of the suspect) and I represents the background information (i.e., the circumstances of the case). The likelihood ratio (LR) is now widely considered the most appropriate framework to report on scientific evidence (Robertson and Vignaux, 1995; Aitken and Taroni, 2004). It provides a measure of the probative value of the finding given the proposition of interest. It is often convenient, due to the wide range of possible values, to convert them to the log-base-ten likelihood ratio. This paper will present a comparison of the log_{10} likelihood ratios obtained using the three different methods to simulated mixtures. Assumptions A1-A5, used for all the three methods, are listed below, while assumptions which are particular to a single method are specified in the corresponding sections.

A1 Each conceptual mixture is composed of the DNA of two contributors. The major contributor's genotype is available and known with certainty. This contributor is referred to as the victim.

A2 The DNA material is in sufficient quantity to obtain all the relevant genotypic information about the contributors that the considered set of markers is supposed to provide (i.e., no allelic drop-out.)

A3 There is no question of a close relative of the suspect being involved.

A4 No DNA artifacts (stutters or drop-in phenomena) occur during the analysis of the mixture.

A5 No subpopulation structures are taken into account.

The idea of the work reported here is to simulate, for each method, $n = 100,000$ mixtures of two contributors, under assumptions **A1** to **A5**, and to calculate the n likelihood ratios both assuming the prosecution's point of view and the defence's point of view. Thus, there is a total of $2n$ likelihood ratios. These values are stored, respectively, in vectors $\mathbf{LR_p}$ and $\mathbf{LR_d}$.

A mixture of two contributors is simulated through the random generation of the four alleles of the contributors, for each locus, with a probability based on the allelic proportions in the population of interest. The prosecution's point of view supposes that the two contributors, referred to as the victim V and the suspect S, are known. When the likelihood ratio for the proposition according to which the suspect is a contributor is calculated for such a mixture, a value greater than one is expected. The higher the likelihood ratio the more interesting is the chosen method from the prosecution's point of view. In this paper the distributions of the likelihood ratio obtained are used to compare the different methods with respect to the prosecution's point of view. Stated otherwise, LR_p is computed for H_p: 'The victim (V) and the suspect (S) contributed to the mixture (i.e., V+S)' and H_d: 'The victim (V) and an unknown person (U) contributed to the mixture (i.e., V+U)', when the mixture E is given by the alleles possessed by V and S.

When the defence's point of view is assumed, a person other than the suspect is considered as a contributor when simulating a mixture. If the suspect's genotype is compatible as a contributor to the mixture, a likelihood ratio higher than one is generally obtained. If the suspect's genotype is not compatible as a contributor to the mixture, a likelihood ratio of 0

is obtained. The higher the number of zero likelihood ratios which are obtained, the more attractive is the method from the defence's point of view. In summary, the defence's point of view is explored by (i) simulating mixtures involving the victim (V) and an unknown contributor (C_2, generated at random), and (ii) calculating likelihood ratios for a target proposition that specifies the suspect (different from C_2 and generated at random) as a second contributor. Again, discrete likelihood ratio distributions are obtained for the different methods, to be used for further comparative analyses. Stated otherwise LR_d is computed for H_p: 'The victim (V) and the suspect (S) contributed to the mixture (i.e., V+S)' and H_d: 'The victim (V) and an unknown person (U) contributed to the mixture (i.e., V+U)', when the mixture E is given by the alleles possessed by V and U. The next section offers details on this.

4.3.1 Likelihood ratios for STR markers

The mixtures which are simulated for the STR markers should all represent moderately unbalanced mixtures, otherwise the use of the STR method would not generally give any evidence of the presence of a second contributor. This means that another assumption should be introduced before evaluating the simulated STR results.

A6 for STR The mixture is moderately unbalanced.

To assess the results obtained from a moderately unbalanced mixture with the standard STR method, the likelihood ratio is calculated, marker by marker, using a Bayesian network proposed in Dawid et al. (2007) and Mortera et al. (2003). The overall likelihood ratio is obtained by the product of the marker specific likelihood ratios, due to the independence assumption made earlier in Section 4.2.2.

In order to simulate STR results for a mixture of two persons under assumptions **A1** to **A5**, four STR alleles (two for each contributor) are drawn, based on the allelic frequencies of the population of interest, for each marker. The first of the two contributors is defined as the victim, while the second is referred to as C_2. When considering the prosecution's point of view, the suspect is assumed to be C_2. Under the defence's point of view, the genotype of a third 'actor', which is the second contributor and is different from the suspect, has to be randomly generated. For each marker, a likelihood ratio is calculated using the Bayesian network, by specifying the alleles of the mixture, the genotype of the suspect and that of the victim. Doing so for all the markers, and multiplying the resulting likelihood ratios, the overall likelihood ratio for each mixture is obtained, depending on the particular point of view assumed (prosecution or defence).

If this process is iterated n times assuming the prosecution's point of view, a vector of n likelihood ratio results, called $\mathbf{LR_p^{STR}}$, is obtained and a discrete distribution for those values can be given. Iterating the process n times, assuming the defence's point of view, the vector $\mathbf{LR_d^{STR}}$ is obtained.

Prosecution's point of view

The $\log_{10}$ likelihood ratios obtained are all extremely high. The minimum value observed for the n simulated mixtures is 13.78, with a mean of 20 (see Table 4.3 for a detailed summary and comparison with the corresponding DIP-STR simulation results). The summaries for the distributions of the $\log_{10}$ likelihood ratios for each of the 16 STR markers are represented in the appendix (Table 4.8). The histogram for this distribution can be inferred from Figure 4.1 (grey bars (a), and grey line (b)).

Defence's point of view

When the defence's point of view is considered, most of the values of $\log_{10}\mathbf{LR_d^{STR}}$ are found to be zero. In fact, while marker specific likelihood ratios are occasionally higher than zero, the likelihood ratios over all markers are all found to be equal to zero (see Table 4.4 for a comparison with the corresponding DIP-STR simulation results). Thus, histograms are not very convenient to present these results, and a tabular summary appears to be more useful. Table 4.9 in the Appendix shows the percentage of values which are equal to 0 or which belong to one of the following intervals: $[1, 10)$, $[10, 100)$, $[100, 1000)$, $[1000, 10,000)$, $> 10,000$. For the interval [0,1] no likelihood ratio values are obtained. This is because, whenever the suspect's genotype is 'compatible' with the profiling results for the trace, the probability of observing the mixture profile given the first proposition (i.e., that the suspect is a contributor) is greater than given the alternative proposition (i.e., that an unknown person unrelated to the suspect is the second contributor). Thus likelihood ratios are either equal to 0, or greater than 1. Note that values greater than one, for this situation, wrongly support hypothesis H_p. For illustration, the bounds of the intervals shown in Table 4.9 are chosen to correspond to those of the scale of likelihood ratios and strength of verbal support in favor of the proposition H_p (Evett et al., 2000).

4.3.2 Likelihood ratios for DIP-STR markers

In Cereda et al. (2014b), an object-oriented Bayesian network (see Appendix, Figure 4.6) was constructed for the assessment of DIP-STR profiling results obtained from a mixture of two contributors (independently of the mixture proportion). This network allows one to obtain the likelihood ratio for the proposition according to which the suspect is the second contributor (versus the proposition that an unknown person is the second contributor), given the assumption that the first contributor is the victim.

The simulation of a mixture of two persons using DIP-STR alleles is similar to the procedure explained in Section 4.3.1. The only difference is that, for a given pair of contributors, possible results consist either of the DIP-STR allele(s) of the minor contributor (if the major contributor is DIP-homozygous and the minor contributor has at least one DIP allele of different kind), or of no alleles (see Table 4.1). As before, two vectors of n likelihood ratios are obtained, denoted here $\mathbf{LR_p^{DIP}}$ and $\mathbf{LR_d^{DIP}}$. Again these can be investigated through their discrete distributions.

Prosecution's point of view

The maximum $\log_{10}$ likelihood ratio observed for the simulated n mixtures is 13.71, which is close to the minimum value observed for the simulations using STR markers. However, for the DIP-STR simulations, the minimum value is 0 and the mean is 3.37 (see Table 4.3 for further summary statistics and a comparison with the results for the STR method). The summaries for the distributions of the $\log_{10}$ likelihood ratios for each of the 9 DIP-STR markers are represented in the Appendix (see Table 4.10). The histogram for this distribution can be inferred from Figure 4.1 (white bars (a), and white line (b)).

Defence's point of view

Table 4.11 in the Appendix represents the percentages of likelihood ratio results that fall into the various categories of probative value. Values equal to zero are obtained for 99.988% of all likelihood ratios (as shown by Table 4.4.)

4.3.3 Likelihood ratios for the Y-STR markers

The method for deriving the likelihood ratio for a mixture using Y-STR markers is different from that used for STR and DIP-STR markers Gill et al. (2001). Due to a lack of recombination, the majority of the Y-chromosome (including all the Y-STR markers currently used in forensic genetics) represents, in effect, a single locus Roewer (2009). Therefore, the independence assumption made for autosomal markers cannot be applied to estimate the population proportion for a Y-STR haplotype.

Moreover, the only situation in which the Y-STR analyses give interesting results is the one in which the major contributor is female and the minor one is male. This is why the following assumption is used for simulating Y-STR results.

A6 for Y-STR The known contributor to the mixture is female while the second, and incriminated one, is a man.

The simulations, in this case, consist in generating n times the Y-STR haplotype of the second contributor and of the suspect, with a probability based on the Y-STR haplotype proportion in the population of interest, and to evaluate the likelihood ratio, following Equation 4.2.

$$LR = \begin{cases} \frac{1}{\gamma_S} & \text{When assuming the prosecution's point of view} \\ \frac{a}{\gamma_{C_2}} & \text{When assuming the defence's point of view} \end{cases} \tag{4.2}$$

where γ_S and γ_{C_2} are, respectively, the population proportions of the Y-STR haplotypes of the suspect and of the actual second contributor to the mixture. Note that they are the same person under the prosecution's hypothesis. The parameter a is 0 every time the two haplotypes are different, otherwise it is 1.

Different approaches are currently available for assessing the rarity of particular Y-STR haplotypes, among which the counting method (Gill et al., 2001; Budowle et al., 2007), the

'haplotype surveying' method Roewer et al. (2000); Krawczak (2001), the k-method Brenner (2010), and the discrete Laplace method Andersen et al. (2013b). A Bayesian method, based on a uniform prior distribution, which is Dirichlet, is retained here to estimate the proportions of different Y-STR hayplotypes in a relevant population. The same method is used for estimating the population proportions of STR and DIP-STR alleles.

Table 4.2 represents the percentage of the different values of $\mathbf{LR_p^{Y-STR}}$ and $\mathbf{LR_d^{Y-STR}}$. Note that here the actual likelihood ratio values are used, instead of the $\log_{10}$, due to the limited extension of the range of values of the two vectors.

LR values	Percentage in $\mathbf{LR_p^{Y-STR}}$	Percentage in $\mathbf{LR_d^{Y-STR}}$
0	0	99.31
58.2	1.67	0.041
72.75	1.39	0.018
97	4.06	0.039
145.5	92.88	0.592

Table 4.2: Percentage of different values of $\mathbf{LR_p^{Y-STR}}$ and $\mathbf{LR_d^{Y-STR}}$.

Note that only four possible distinct likelihood ratio values (different from 0) are obtained. This is due to the fact that in the considered database (Haas et al., 2006), there are 4 different haplotypes which appear twice (and thus bring to a likelihood ratio of 97), one haplotype which appears three times (likelihood ratio of 72.75), one which appears four times (likelihood ratio of 58.2) while the other 135 different haplotypes appear only once each (likelihood ratio of 145.5). Likelihood ratios equal to zero are obtained, when using simulation for the defence point of view, each time the Y-STR genotype of the second contributor and of the suspect are not compatible.

The use of the assumption **A3** about the impossibility of kinship between the perpetrator and the suspect under the hypothesis H_d has a strong effect on the likelihood ratio values for Y-STR profiling results: in fact, as noted earlier, if no mutations occur, patrilineal relatives of the suspect share the same Y-STR profile, which would imply different values for the likelihood ratio if one takes them into account.

4.4 Comparison of the three methods

This section compares the DIP-STR method with both the regularly used STR method, and the Y-STR method. The comparison is based on the distributions of the likelihood ratio results obtained with the three methods, assuming the same point of view (i.e., of the prosecution, or of the defence).

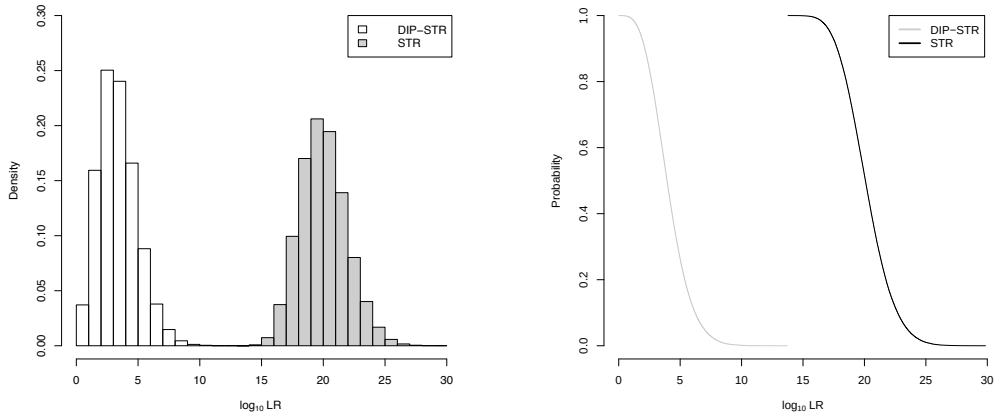
The comparison of the DIP-STR and the STR likelihood ratio results supposes moderately unbalanced mixtures because, otherwise, the use of STR markers is likely to miss any indication of the presence of a second person in the mixture. The comparison between the

DIP-STR and Y-STR likelihood ratio results assumes mixtures which could involve any unbalance proportion, but with the constraint that the major contributor is a women and the minor one is a man. Note that the latter comparison becomes relevant in case of extremely unbalanced mixtures, that is when STR markers can generally not be used.

The comparisons from the prosecution point of view are carried out by plotting in the same graph the histograms of the distributions of $\log_{10}\mathbf{LR_p}$ for the two methods, and in another graph their Tippett plots. The latter are graphical representations first reported for forensic DNA evaluation in Evett and Buckleton (1996), and inspired by the concepts of 'within-source comparison' and 'between-sources comparison' as defined by Tippett Tippett et al. (1968). In this kind of representation, the x axis represents the different ($\log_{10}$) values of the likelihood ratio from the prosecution point of view. The y axis represents the proportion of cases in which the likelihood ratio exceeds the corresponding value in the x axis.

4.4.1 Comparison of DIP-STR and STR assuming point of view of the prosecution

Before the comparison is performed in further detail, it is worth recalling that this is meaningful only under the assumption **A6 for STR**, that is in case of moderately unbalanced mixtures. Figure 4.1 represents the histograms and the Tippets plots for $\log_{10}$ of the likelihood ratio values for the two methods, assuming the prosecution point of view. Table 4.3 presents the standard summary statistics for the two distributions.



(a) (b)

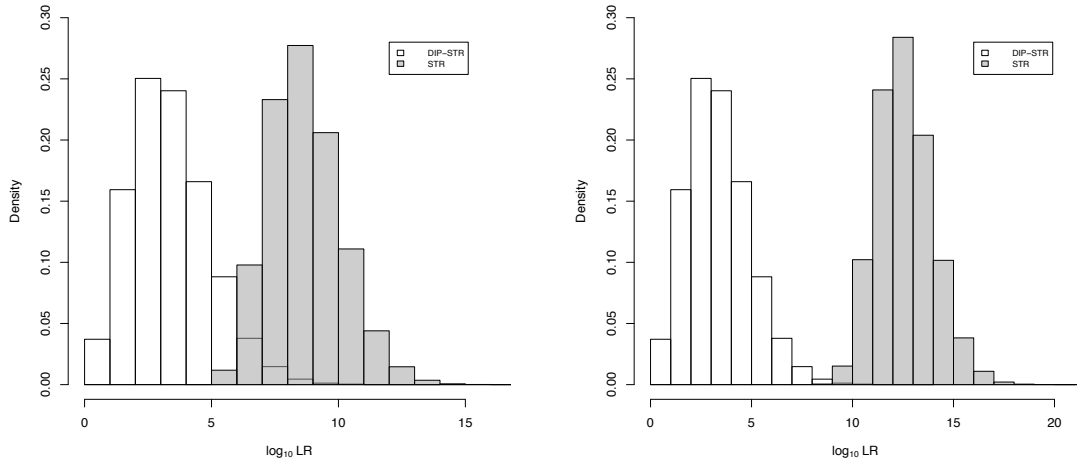
Figure 4.1: Graphical comparisons of the $\log_{10}\mathbf{LR_p^{STR}}$ and $\log_{10}\mathbf{LR_p^{DIP}}$ distributions in terms of superimposed histograms (a) and Tippett plots (b).

Figure 4.1 shows that the distribution of the likelihood ratio values for the STR markers is shifted towards higher values than the distribution of the likelihood ratio for the DIP-STR markers. This means that, from the prosecution's point of view, the use of the STR kit is more desirable. It has to be noticed, however, that since the STR kit has 7 markers more than the DIP-STR kit, this difference is little surprising.

Marker system	Min	1st Quantile	Median	Mean	3rd Quantile	Max
DIP-STR	0	2.228	3.201	3.367	4.324	13.706
STR	13.781	18.658	19.899	19.996	21.218	29.85

Table 4.3: The summaries of the distributions of $\log_{10}\mathbf{LR_p^{STR}}$ and $\log_{10}\mathbf{LR_p^{DIP}}$.

In order to arrange a comparison using the same number of markers for the two methods, 9 STR markers were chosen here out of the 16. There are 11,440 combinations of 9 markers out of 16, but here we have focused on the two combinations of 9 markers for which the means of the distributions of the $\mathbf{LR_p^{STR}}$ are, respectively, minimally and maximally separated of the distribution of the $\mathbf{LR_p^{DIP}}$. These two combinations have been found empirically, running simulations for each of combinations. Figure 4.2 shows the histograms for the distribution of $\log_{10}\mathbf{LR_p}$ for the two methods, using these two combinations of 9 out of 16 STR markers in comparison with the histogram for the distribution of $\log_{10}\mathbf{LR_p^{DIP}}$. Table 4.4 shows the summaries for the 3 distributions. These results confirm the previous finding: the STR markers system performs better than the DIP-STR marker system.



(a) (b)

Figure 4.2: Comparisons of the distribution of $\log_{10}\mathbf{LR_p^{DIP}}$ (white bars) with the distribution of $\log_{10}\mathbf{LR_p^{STR}}$ (grey bars) for the combination of markers for which the mean of the two distributions are (a) minimally and (b) maximally separated.

Marker system	Min	1st Quantile	Median	Mean	3rd Quantile	Max
DIP-STR	0	2.228	3.201	3.367	4.324	13.706
STR (min. separated)	4.800	7.653	8.557	8.677	9.578	16.140
STR (max. separated)	8.905	12.530	13.500	13.600	14.560	21.620

Table 4.4: Summaries of the distribution of $\log_{10}\mathbf{LR_p^{DIP}}$ and $\log_{10}\mathbf{LR_p^{STR}}$ choosing the 9 STR markers for which the means of the distributions are the minimally (second row) and the maximally (third line) separated.

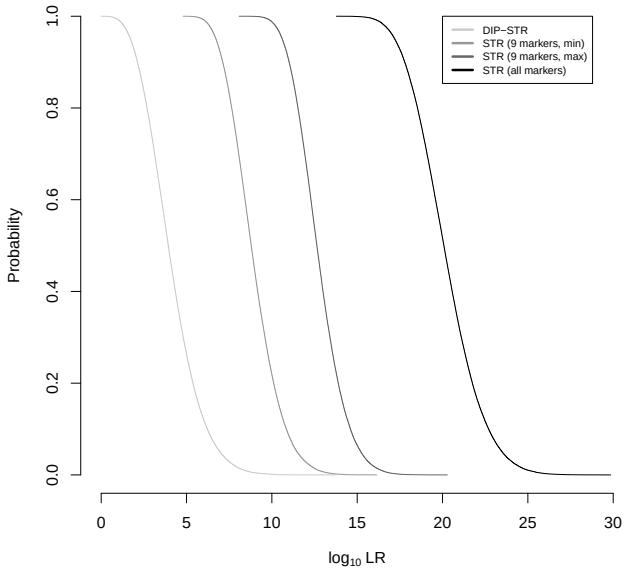


Figure 4.3: Tipett plots of the $\log_{10}\mathbf{LR}_{\mathbf{p}}^{\mathbf{DIP}}$ and $\log_{10}\mathbf{LR}_{\mathbf{p}}^{\mathbf{STR}}$ results for both, full STR profiles and profiles with reduced numbers of markers.

Figure 4.3 shows the Tippett plots of the DIP-STR $\log_{10}$ likelihood ratio distribution and the 3 different distributions of STR $\log_{10}$ likelihood ratios (i.e., one for full STR profiles, and two with only 9 markers). As may be seen, the two likelihood ratio distributions with 9 STR markers are closer to the DIP-STR likelihood ratio distribution than the one with 16, just as expected.

4.4.2 Comparison between DIP-STR and STR marker systems assuming the point of view of the defence

Tables 4.4 summarises the percentage of values of $\mathbf{LR}_{\mathbf{d}}^{\mathbf{STR}}$ and $\mathbf{LR}_{\mathbf{d}}^{\mathbf{DIP}}$ that fall into the different intervals of likelihood ratio values, corresponding to different expressions of probative strength.

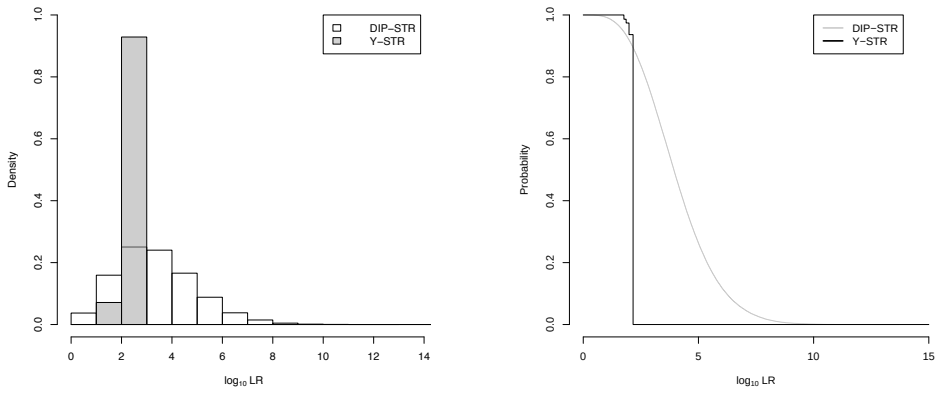
LR_d	Verbal equivalent	DIP-STR markers	STR markers
0	Exclusion	99.988	100
1-10	Limited support	0	0
10-100	Moderate support	0	0
100-1000	Moderately strong support	0.003	0
1000-10,000	Strong support	0.007	0
$> 10,000$	Very strong support	0.002	0

Figure 4.4: Percentage of DIP-STR and STR likelihood ratio values found for various intervals of probative strength for the hypothesis H_d.

From the defence's point of view the more the number of zeros among the simulated likelihood ratios, the more the method is desirable. Hence, Table 4.4 indicates that from the defence's point of view there is an advantage in using STR markers (for balanced mixtures), because the proportion of likelihood ratio values with 0 is maximal, while using DIP-STR markers 0.012% of simulated cases offer a false positive. In principle, the same considerations outlined in Section 4.4.1, which ascribe the difference in the overall likelihood ratio distribution to the different number of markers in the two kits, can be made in the case here. But even if one chooses the 9 STR markers which have the highest number of non-zero values (D8, D3, D1S, D12, VWA, D2S1, D18, FGA, D2S4) and then multiply them to obtain the overall likelihood ratio, one comes to the same conclusion, since 100% of 0 likelihood ratio values are obtained.

4.4.3 Comparison between DIP-STR and Y-STR marker systems assuming the point of view of the prosecution

Before proceeding with the details of the comparison between DIP-STR and Y-STR, it is useful to recall that this comparison is meaningful only under assumption **A6 for Y-STR**, that is when the known contributor is female and the unknown is a male. No assumption is needed, however, about the mixture proportion. As in Sections 4.4.1 and 4.4.2, the two likelihood ratio distributions to be compared are represented in terms of superimposed histograms and Tippett plots in Figure 4.5.



(a) (b)

Figure 4.5: Comparisons of the $\log_{10}\mathbf{LR_p^{Y-STR}}$ and $\log_{10}\mathbf{LR_p^{DIP}}$ distributions using superimposed histograms (a) and Tippett plots (b).

Since the histogram for the distribution of Y-STR likelihood ratio is composed by only two bars, Table 4.5 is retained here as a tabular summary.

Table 4.5 and Figure 4.5 indicate that, from the point of view of the prosecution, the use of DIP-STR markers appears more useful than that of Y-STR markers. With the latter, one can obtain at best a moderately strong support, while with DIP-STR markers an equal or

LR_p	Verbal equivalent	DIP-STR markers	Y-STR markers
0	Exclusion	0	0
$1-10$	Limited support	3.708	0
$10-100$	Moderate support	15.939	7.123
$100-1000$	Moderately strong support	25.037	92.877
$1000-10,000$	Strong support	24.026	0
$>10,000$	Very strong support	31.288	0

Table 4.5: Percentage of $\mathbf{LR_p^{DIP}}$ and $\mathbf{LR_p^{Y-STR}}$ values which fall into different categories of probative strength for H_p

higher degree of support is obtained in more than 80% of the cases, while a lower degree of support is obtained only in less than the 4% of the cases.

4.4.4 Comparison between DIP-STR and Y-STR marker systems assuming the point of view of the defence

Table 4.6 summarises the percentage of likelihood ratio values that fall into different categories of probative value, for simulations performed according to the viewpoint of the defence.

LR_d	Verbal equivalent	DIP-STR markers	Y-STR markers
0	Exclusion	99.988	99.31
$1-10$	Limited support	0	0
$10-100$	Moderate support	0	0.098
$100-1000$	Moderately strong support	0.003	0.592
$1000-10,000$	Strong support	0.007	0
$>10,000$	Very strong support	0.002	0

Table 4.6: Percentage of $\mathbf{LR_d^{DIP}}$ and $\mathbf{LR_d^{Y-STR}}$ values that fall into different categories of probative strength for H_p.

This table indicates that a comparison between DIP-STR and Y-STR markers (from the point of view of the defence), should take into consideration two factors. First, if one seeks to be conservative about the number of times in which a likelihood ratio greater than zero is obtained, the use of DIP-STR markers appears slightly preferable. Second, if one seeks to be conservative with respect to the strength of support obtained for values which are greater than zero, then Y-STR markers should be preferred, since – at worst – only a moderate support is obtained for those cases. Stated otherwise, one can consider two main options. One, the DIP-STR method, involves a higher number of zero likelihood ratio values, but with some possibility of a high likelihood ratio against a suspect who has a genotype 'compatible' with a mixture to which he is *not* a contributor. The other, the Y-STR method, involves a higher rate of likelihood ratios that would wrongly associate a suspect with a mixture. However, the likelihood ratios for such cases would be more moderate than in case of the DIP-STR method.

4.4.5 A discussion about the influence of genetic model assumptions

The problem of estimating Y-STR haplotype proportions is a fundamental one Brenner (2010). As already explained in Section 4.3.3, a Bayesian method is retained here for overall consistency with respect to what has been done for the STR and the DIP-STR methods. It is worth to emphasize, however, that the choice of a different method can lead to different simulation results and, consequently, to different conclusions about the comparison between the DIP-STR and the Y-STR methods. Among the alternative methods, the k-method Brenner (2010) and the discrete Laplace method Andersen et al. (2013b) have been chosen to investigate how substantial the difference in the conclusions would be. The k-method leads to essentially the same conclusions as those described above, both for the prosecution's and the defence's point of view. The choice of the discrete Laplace method results in a substantially different distribution for $\mathbf{LR_p^{Y-STR}}$, which would make the Y-STR method preferable from the prosecution's point of view. From the defence's point of view, the use of this method makes the DIP-STR and the Y-STR methods almost equivalent. This points out that there is a strong dependency on population genetic model assumptions. It is worthy of emphasis that there are inherent limitations in the state of the art, and whatever method is applied, its strengths and limitations should be carefully considered.

4.5 Consideration on the usefulness of the three methods

This section pursues a discussion on the proportion of cases in which each of the three methods cannot be used and therefore gives useful input to decision makers on their choice of the analytical methodology.

With regards to the STR method, it has already been explained that, in case of extremely unbalanced mixtures, this method is generally not useful to detect the minor contributor (see Section 4.1). In current practice, many or most extremely unbalanced mixtures probably go undetected, so that it appears difficult to assess the proportion of cases in which such mixtures are encountered.

In turn, it is easier to circumscribe the proportion of cases in which Y-STR markers are not useable. As noted in Section 4.2.3, that is the case whenever the major and the minor contributors are not a female and a male, respectively.

With regards to DIP-STR markers, there is only one situation in which this marker system is not useful. That is, when for all nine DIP-STR markers the major contributor is DIP-heterozygous (see also Table 4.1). In fact, as explained in Section 4.2.1, in case the known contributor is homozygous for the DIP allele, the fact of obtaining no alleles for the second contributor gives information about the DIP alleles of the minor contributor. The proportion of such kind of cases in the population can be assessed using the estimated allele proportions of each marker. This result is displayed in Table 4.7, which provides, for each marker, the probability that an individual (taken here as the major contributor) is heterozygous, or that both contributors are homozygous for same DIP allele (S or L), within the corresponding

likelihood ratio. Actually, these are cases in which the likelihood ratio has the lowest values, independently on the STR parts which constitute the DIP-STR minor contributor's genotype. The last column in the table gives the probability that in all markers the major contributor is DIP-heterozygous, or that both contributors are homozygous for the same DIP alleles (S or L).

	Marker 1		Marker 2		Marker 3	
	LR	Probability	LR	Probability	LR	Probability
Major heterozygous	1	0.374	1	0.475	1	0.442
Both homozygous S	1.772	0.318	2.678	0.139	9.17	0.012
Both homozygous L	16.146	0.004	6.612	0.023	2.229	0.201

	Marker 4		Marker 5		Marker 6	
	LR	Probability	LR	Probability	LR	Probability
Major heterozygous	1	0.335	1	0.475	1	0.342
Both homozygous S	1.613	0.384	2.678	0.139	20.816	0.002
Both homozygous L	22.11	0.002	6.612	0.0229	1.640	0.372

	Marker 7		Marker 8		Marker 9	
	LR	Probability	LR	Probability	LR	Probability
Major heterozygous	1	0.457	1	0.403	1	0.077
Both homozygous S	2.392	0.175	1.931	0.268	1.086	0.848
Both homozygous L	8.003	0.016	12.721	0.006	613.938	2.65×10^{-6}

	All markers	
	LR	Probability
Major heterozygous	1	6.12×10^{-5}
Both homozygous S	19631.581	2.59×10^{-9}
Both homozygous L	$3{,}57\times10^{9}$	7.86×10^{-20}

Table 4.7: The probability of occurrence of the 3 lowest likelihood ratio values obtained with DIP-STR markers, namely the values corresponding to cases in which the major contributor is heterozygous, or both contributors are homozygous for the same DIP allele.

It is worth noting that probabilities in Table 4.7 are not derived from the simulations of mixtures. They are calculated on the basis of the allele proportions relating to the databases of interest. The probability of a genotype that in all markers is heterozygous for the DIP allele is 6.12×10^{-5} (see last column of Table 4.7). This means that, on average, in only about 0.00612% of the cases a mixture, analysed with DIP-STR markers, does not help in discriminating between the two hypotheses of interest. This proportion seems remarkably small. In an actual case, it may thus be of interest to compare this proportion with the probability of facing an unbalanced mixture that may not lead to appropriate results with the traditional STR technique (to be assessed in the light of the case circumstances). However, this argumentation takes as an assumption that the mixture has already been recognised as such. In fact, there are situations (typically the case in which the two contributors are DIP-homozygous of the same type) in which a LR $\neq 1$ is obtained (as already pointed out in Section 4.2.1, only if one presumes the presence of a second contributor and the genotype of a suspect is also available.

Contrary to what happens with the use of STR markers and Y-STR markers, where often the second contributor to the mixture is missed without even suspecting his (her) presence,[2] with the use of DIP-STR markers it is sometimes possible to know with certainty and in advance the impossibility of detecting the genotype of the second contributor (i.e., when the major contributor is DIP-heterozygous in all markers). In general, using DIP-STR markers a mixture cannot be recognised as such when in all markers either the major contributor is heterozygous or both contributors are DIP-homozygous of the same type. The probability that an actual two-person mixture will not be recognised as such (i.e., the presence of a second contributor cannot be pointed out) has been calculated, using the allelic proportions, as the probability that in each marker either the major contributor is DIP-heterozygous or the two contributors are DIP homozygous of the same type. This probability is equal to 0.039. This means that about 4% of recovered stains, which are actually mixtures, will leave one with uncertainty about the presence of a second contributor.

4.6 Conclusion

The research reported in this article aimed at comparing three profiling methods for analysing DNA mixtures of two contributors. The relative advantages and limitations of STR markers, DIP-STR markers and Y-STR markers was considered from the point of view of the defence and the prosecution. In such a comparison, different aspects appear relevant, such as the proportion of cases in which mixtures have characteristics that make a given method useful (see, e.g., Section 4.5), and the distribution of likelihood ratio results in scenarios that reflect the viewpoint of either the prosecution or the defence (i.e., propositions of interest H_p and H_d, as defined in Section 4.3).

For cases of, at worst, moderately unbalanced mixtures, the simulation results − that is the distributions of the likelihood ratio values both from the prosecution's and the defence's point of view − suggest that the traditional STR marker system should be preferred. The case is different for extremely unbalanced mixtures. Here, STR markers are not reliable, but Y-STR markers and DIP-STR markers are applicable (Section 4.5). In such cases, the latter method should be preferred from the prosecution's point of view, since in about the 80% of the cases one obtains likelihood ratios which are higher than those obtained with the Y-STR method. However, from the the defence's point of view, two aspects should be reminded: one aspect concerns the strength of support obtained in case of a wrong association (i.e., when the likelihood ratio supports the wrong proposition), the other aspect relates to the number of times in which such a wrong indication is encountered. This is why from the defence's point of view, preferences may depend on what aspect one considers.

The common way to detect the presence of a possible second contributor to a stain already typed for STR markers (and which appeared as a single mixture), is to use Y-STR markers. However, this approach too, can miss the minor contributor if the gender composition of the two contributors is not proper (i.e., the major one is female and the minor one is male). The

[2]With the use of quantification methods it is possible to detect the presence of a second contributor, but only for the good gender mismatch between the two contributors: the major one should be female and the minor one should be male.

use of DIP-STR markers can thus be desirable for all those kind of traces that, with the use of STR and Y-STR markers, appear as single stain, but for which one suspects the presence of a second contributor. In these cases, DIP-STR markers can also complement Y-STR results to discriminate paternally related individuals.

Actually, the use of DIP-STR markers could present an interest for all kind of DNA stains, independently of the use of STR markers. The reason for this is that with the use of DIP-STR markers one can establish in advance if this method could be used, because it starts by determining the genotype of the assumed known major contributor (see Section 4.5). In case of a favourable outset, DIP-STR profiling can provide information about the second contributor in terms of one, two or no alleles (Section 4.2.1). Although the likelihood ratio distributions obtained under the defence's and the prosecution's point of view are not as marked as those that can be obtained with traditional STR markers, they can still be regarded as practically useful (see, e.g., Tables 4.10 and 4.11). In addition, new DIP-STR markers are currently investigated. This may favourably improve the likelihood ratio distributions that could be obtained under the various competing points of view in a near future, but analysts should also remind that the definition of practical procedures will also encompass additional factors such as time and monetary constraints.

Appendix A. Additional tables and figures

Marker name	Min	1st Quantile	Median	Mean	3rd Quantile	Max
VWA	0.54	0.74	0.97	1.04	1.27	2.36
TH01	0.45	0.62	0.8	0.86	0.99	2.57
SE33	1.48	2.04	2.29	2.38	2.65	4.37
FGA	0.78	0.98	1.269	1.35	1.67	4.33
D22	0.37	0.46	0.68	0.86	0.99	4.32
D21	0.59	0.85	1.13	1.27	1.5	4.33
D19	0.42	0.62	0.94	1.14	1.54	4.33
D18	0.89	1.13	1.39	1.48	1.69	3.93
D16	0.43	0.58	0.78	0.92	1.1	3.77
D12	0.98	1.35	1.63	1.71	2.06	4.34
D10	0.43	0.58	0.76	0.89	1.08	3.09
D8	0.51	0.77	1.07	1.14	1.37	4.32
D3	0.56	0.66	0.80	0.87	1.03	2.70
D2S4	0.34	0.47	0.77	0.94	1.35	4.15
D2S1	0.78	1.18	1.43	1.50	1.73	4.33
D1S	0.97	1.36	1.58	1.63	1.85	4.16
All markers	13.78	18.66	19.9	20	21.22	29.85

Table 4.8: The summaries of the distributions of the $\log_{10}$ of the likelihood ratio values for each STR marker and for the overall $\log_{10}\mathbf{LR}_{\mathbf{p}}^{\mathbf{STR}}$ (last row).

LR	Verbal equivalent	VWA	TH01	SE33	FGA	D22	D21
0	Exclusion	92.191	89.122	99.464	95.572	85.816	93.698
$1-10$	Limited support	5.818	9.694	0	1.889	13.415	4.377
$10-100$	Moderate support	1.981	1.175	0.221	2.513	0.749	1.852
$100-1000$	Moderately strong support	0.01	0.009	0.306	0.025	0.02	0.07
$1000-10,000$	Strong support	0	0	0.009	0.001	0	0.003
$>10,000$	Very strong support	0	0	0	0	0	0

LR	Verbal equivalent	D19	D18	D16	D12	D10	D8
0	Exclusion	90.169	96.695	88.581	97.85	88.229	92.437
$1-10$	Limited support	8.593	0.606	9.478	0.05	10.659	5.41
$10-100$	Moderate support	1.181	2.645	1.904	1.963	1.102	2.123
$100-1000$	Moderately strong support	0.057	0.053	0.021	0.135	0.008	0.029
$1000-10,000$	Strong support	0	0.001	0	0.002	0.002	0.001
$>10,000$	Very strong support	0	0	0	0	0	0

LR	Verbal equivalent	D3	D2S4	D2S1	D1S	All markers
0	Exclusion	90.102	86.229	96.701	97.682	100
1-10	Limited support	8.116	12.735	0.937	0.023	0
10-100	Moderate support	1.775	1.01	2.293	2.244	0
100-1000	Moderately strong support	0.007	0.026	0.068	0.051	0
1000-10,000	Strong suppor	t0	0	0.001	0	0
$>10,000$	Very strong support	0	0	0	0	0

Table 4.9: Percentage of likelihood ratio values belonging to the different intervals of probative value in favour of the proposition H_p, for each STR marker and combined across all markers (last column). Marker names are abbreviated to their first three characters.

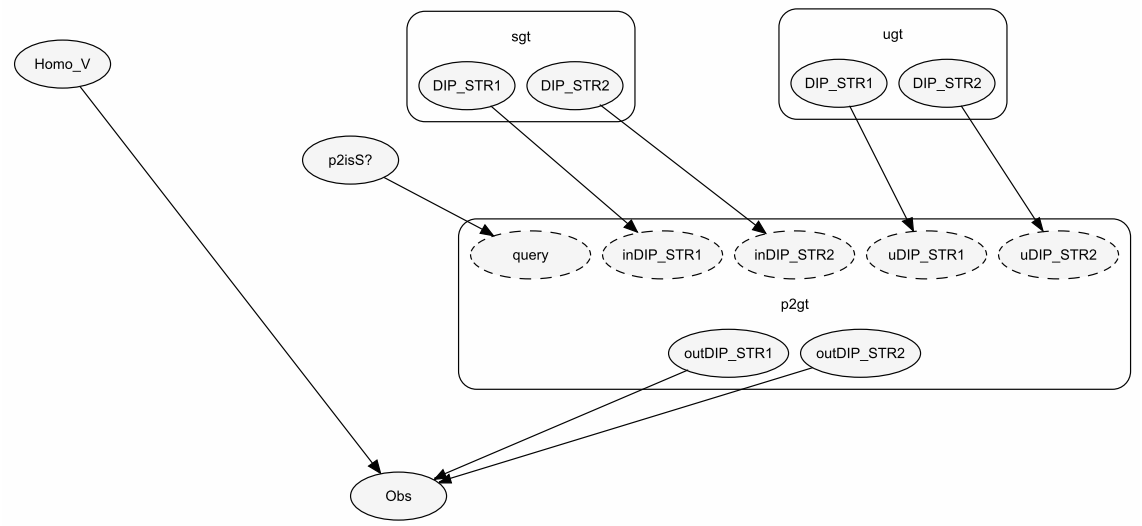


Figure 4.6: Object-oriented Bayesian network for evaluating DIP-STR profiling results of mixtures from two contributors, when DIP-STR markers are used (Cereda et al., 2014b).

Marker name	Min	1st Quantile	Median	Mean	3rd Quantile	Max
Marker 1	0	0	0.25	0.46	0.94	3.79
Marker 2	0	0	0.43	0.41	0.60	3.77
Marker 3	0	0	0.35	0.41	0.52	3.76
Marker 4	0	0	0.21	0.41	0.90	3.61
Marker 5	0	0	0.43	0.44	0.71	3.59
Marker 6	0	0	0.21	0.35	0.78	3.78
Marker 7	0	0	0.24	0.37	0.63	3.46
Marker 8	0	0	0.29	0.40	0.74	3.58
Marker 9	0	0.04	0.04	0.11	0.04	3.62
All markers	0	2.23	3.20	3.37	4.32	13.71

Table 4.10: The summaries of the distributions of the $\log_{10}$ of the likelihood ratio values for each DIP-STR marker and for the overall $\log_{10}\mathbf{LR}_{\mathbf{p}}^{\mathbf{DIP}}$ (last row).

LR	Verbal equivalent	Marker 1	Marker 2	Marker 3	Marker 4	Marker 5
0	Exclusion	63.353	77.401	70.57	58.877	79.069
$1-10$	Limited support	35.317	21.579	28.179	40.314	19.924
$10-100$	Moderate support	1.312	0.997	1.244	0.755	0.997
$100-1000$	Moderately strong support	0.017	0.023	0.007	0.054	0.01
$1000-10,000$	Strong support	0.001	0	0	0	0
$>10,000$	Very strong support	0	0	0	0	0

LR	Verbal equivalent	Marker 6	Marker 7	Marker 8	Marker 9	All markers
0	Exclusion	58.486	69.142	67.975	14.802	99.988
$1-10$	Limited support	41.272	29.543	31.262	84.548	0
$10-100$	Moderate support	0.239	1.223	0.753	0.65	0
$100-1000$	Moderately strong support	0.003	0.092	0.01	0	0.003
$1000-10,000$	Strong support	0	0	0	0	0.007
$>10,000$	Very strong support	0	0	0	0	0.002

Table 4.11: Percentage of likelihood ratio values obtained for the various categories of probative value in favour of H_p, for each marker and combined across all markers, when the defence's point of view is considered.

Chapter 5

Impact of model choice on LR assessment in case of rare haplotype match (frequentist approach)

This chapter is based on:
Cereda, G. (2016) Impact of model choice on LR assessment in case of rare haplotype match (frequentist approach). *Scandinavian Journal of Statistics*, In Press.

Abstract

The likelihood ratio (LR) measures the relative weight of forensic data regarding two hypotheses. Several levels of uncertainty arise if frequentist methods are chosen for its assessment: the assumed population model only approximates the true one and its parameters are estimated through a database. Moreover, it may be wise to discard part of data, especially that only indirectly related to the hypotheses. Different reductions define different LRs. Therefore, it is more sensible to talk about "a" LR instead of "the" LR, and the error involved in the estimation should be quantified. Two frequentist methods are proposed in the light of these points for the 'rare type match problem', that is when a match between the perpetrator's and the suspect's DNA profile, never observed before in the database of reference, is to be evaluated.

5.1 Introduction

One of the main challenges of forensic science is to evaluate how much some evidence can be helpful to discriminate between hypotheses of interest. For instance, a typical piece of evidence may be a DNA trace which is found at the crime scene and whose profile matches a known suspect's DNA profile. A couple of mutually exclusive hypotheses is typically defined, of the kind of 'the crime stain came from the suspect' (h_p) and 'the crime stain came from an unknown donor' (h_d). The largely accepted method to perform this evaluation is the

calculation of the *likelihood ratio*, a statistic that expresses the relative plausibility of the observations under the two hypotheses (Robertson and Vignaux, 1995; Evett and Weir, 1998; Aitken and Taroni, 2004; Balding, 2005; Steele and Balding, 2014).

The definition of the likelihood ratio depends on whether a Bayesian or a frequentist approach is chosen. In the Bayesian context, after a couple of hypotheses is given, the likelihood ratio is defined as

$$\mathrm{LR} = \frac{\Pr(D = d \mid H = h_p)}{\Pr(D = d \mid H = h_d)}, \tag{5.1}$$

where Pr is the Bayesian probability, reflecting the expert's belief on the joint distribution of the random variables of the model, namely D (representing the data), H (representing the hypotheses), and Θ (a nuisance parameter(s)).

On the other hand, in a frequentist context, the nuisance parameter θ and the hypotheses h are considered to be fixed (unknown) quantities. The frequentist probability (here denoted as $\mathscr{Pr}$) can be expressed in terms of the Bayesian Pr, in the following way: $\mathscr{Pr}_\theta(\cdot \mid h) := \Pr(\cdot \mid \Theta = \theta, H = h),\ \forall h$. The frequentist likelihood ratio can be thus expressed as

$$\mathscr{LR}_\theta = \frac{\mathscr{Pr}_\theta(D = d \mid h_p)}{\mathscr{Pr}_\theta(D = d \mid h_d)}. \tag{5.2}$$

It is important to consider that different reductions of the data D can be carried out, each corresponding to a different frequentist likelihood ratio. Moreover, unless we choose non-parametric solutions, a model choice is also performed, and there are often parameters to be estimated. Hence, two further levels of uncertainty have to be added to the initial uncertainty regarding which hypothesis is the true one.

The main aim of this paper is to provide the message that, if a frequentist approach is chosen and estimation is needed, (i) it is more sensible to talk about "a" likelihood ratio instead of "the" likelihood ratio, and (ii) a quantification of the error involved in the estimation of the likelihood ratio is to be provided along with the estimated value.

It is believed in the forensic field that the use of frequentist methods to assess the likelihood ratio is not coherent, since the likelihood ratio has to be used within the Bayes' theorem context, as the way to update prior odds to posterior odds. However, frequentists may be interested as well in the likelihood ratio, seen as a tool to measure the evidential value of data, independently of the Bayes' theorem. Moreover, literature presents many approaches to calculate the likelihood ratio, wrongly defined as Bayesian, which in fact plug in Bayes estimates into a likelihood ratio defined in a frequentist way (for a discussion, see Cereda, 2016a). We thus believed that it is important to study and discuss the two approaches (the Bayesian and the frequentist) separately, in order to define coherent methodologies and avoid unnecessary hybrid methods. This is done in Section 5.2.

In forensic science, a very challenging problem is the so-called *rare type match*, the situation in which there is a match between the characteristics of some recovered material and the corresponding characteristics of the control material, but these characteristics have not been observed yet in previously collected samples (i.e., they do not occur in any existing database of interest for the case). This constitutes a problem because of the presence of a nuisance

parameter that is (related to) the proportion of individuals (or items) in possess of the matching characteristic in a reference population: this proportion is, in standard frequentist practice, estimated using the relative frequency of the characteristic in a previously collected database. Thus, in case of rare type match there's the need for different solutions.

This paper discusses two frequentist methods to provide a likelihood ratio in the rare type match case, based respectively on the parametric Discrete Laplace method (Andersen et al., 2013b), and on a generalization of the nonparametric Good-Turing estimator (Good, 1953). The latter looks similar to Brenner's 'κ-method' (Brenner, 2010), but is different inasmuch it does not need any assumption and provides two different frequencies, one for the prosecution's and one for the defense's point of view. We plan to compare the two methods in a future paper.

More specifically, these two methods are here proposed as an answer to the problem of the rare Y-STR haplotype match: the situation in which the matching (and previously unseen) characteristic is a Y-STR profile. Each of the two methods is analysed in the light of points (i) and (ii) discussed above, by carefully specifying the data reduction, the chosen probability model, and with a discussion on the different levels of error involved in the estimations.

Sections 5.3 and 5.4 draw out in depth the rationale behind points (i) and (ii) above, Section 5.5 describes the paradigmatic example of the rare Y-STR haplotype match problem, to which we will apply the Discrete Laplace method (Section 5.6), and the Generalized Good method (Section 5.7) according to the guidelines exposed in the opening sections.

5.2 Bayesian versus frequentist approach to likelihood ratio assessment

The task of a forensic statistician is to measure the extent to which some given data favors one hypothesis instead of the other. For instance, the data at disposal may consist of a DNA trace found at the crime scene which matches a suspect's DNA profile, and of a database of collected DNA profiles from a reference population or past cases. This is a paradigmatic example to which, from now on, we will refer generically as "the DNA example". The prosecution and defence hypotheses are usually of the kind "the trace has been left by the suspect" (h_p) and "the trace has been left by an unknown person" (h_d). Denote with $h \in \{h_d, h_p\}$ the unknown true hypothesis, and with θ the nuisance parameter involved in the assessment of the likelihood ratio. In the DNA example, the vector made of all the DNA frequencies can be thought of as the nuisance parameter θ. Notice that there is a difference between h and θ: one (h) is the parameter which we 'test' through the likelihood ratio, the other (θ) is a nuisance parameter involved in the likelihood ratio assessment. It is often possible to split the data D into E, evidence directly related to the crime, and B, additional information not related to the crime and only pertaining to the nuisance parameter θ. In the DNA example, we can take as E the couple of matching profiles (that of the trace and that of the suspect) and as B the database of reference. D, E, and B can be regarded as random variables, such that $D = (E, B)$.

Bayesian and frequentist methods differ in how they consider the parameters θ and h. In

a Bayesian context they are modelled through random variables Θ and H, which are given prior distributions $p(\theta)$ and $p(h)$. Frequentists consider them as fixed (i.e., without distribution) unknown quantities. Regardless of the type of approach which is chosen, some model assumptions concerning E and B, θ and h can be made:

a. The distribution of B given h and θ, only depends on θ.

b. B is independent of E, given h and θ.

In the DNA example, condition **a** holds if for instance the database is collected before the crime, since the sampling mechanism to obtain the database of reference is independent of which hypothesis is correct. Condition **b** holds if the suspect has been found on the ground of different evidence that has nothing to do with DNA.

5.2.1 The Bayesian approach

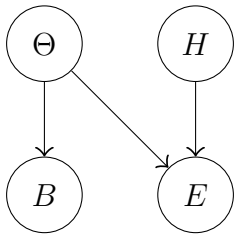


Figure 5.1: Bayesian network representing the dependency relations between E (evidence of the case) B (background data) Θ (nuisance parameter) and H (hypotheses of interest).

A full Bayesian model is defined by giving the prior joint probability distribution for all the random variables of the model (here E, B, H and Θ). It can be represented by the Bayesian network of Figure 5.1, which is in turn equivalent to the following Bayesian reformulation of conditions **a**, and **b**, with a third additional condition:

Bayesian a. B is conditionally independent of H given Θ.

Bayesian b. B is conditionally independent of E given Θ and H.

Bayesian c. Θ is unconditionally independent of H.

Condition **Bayesian c** is guaranteed for instance if prior beliefs on θ and on h are assessed by people with different responsibilities and tasks: a judge for h and a forensic DNA expert (or a statistician) for θ. The joint prior can be factorized as follows, by looking at the structure of the Bayesian network or, equivalently, using the three conditions above: $p(\theta, h, b, e) = p(\theta)p(h)p(b|\theta)p(e|\theta, h)$. By choosing a prior distribution for θ and h which reflects expert's beliefs, the Bayesian probability is an expression of the subjective belief of the experts. The distribution of all other variables given θ and h is defined by the structure of the model, and needs no subjective assessment.

The Bayesian likelihood ratio can be derived in the following way:

$$\begin{aligned}\mathrm{LR} =& \frac{\Pr(E=e, B=b \mid H=h_p)}{\Pr(E=e, B=e \mid H=h_d)} = \frac{\Pr(E=e \mid B=b, H=h_p)}{\Pr(E=e \mid B=b, H=h_d)} = \frac{\int p(e \mid b, h_p, \theta)\, p(\theta \mid b, h_p) \mathrm{d}\theta}{\int p(e \mid b, h_d, \theta) p(\theta \mid b, h_d) \mathrm{d}\theta} \\ =& \frac{\int \theta\, p(\theta \mid b)\, \mathrm{d}\theta}{\int \theta^2\, p(\theta \mid b)\, \mathrm{d}\theta} = \frac{\mathbb{E}(\Theta \mid B=b)}{\mathbb{E}(\Theta^2 \mid B=b)}.\end{aligned}$$

Some simplifications have been carried out because of conditions **a**, **b**, and **c**. Since it is possible to marginalize out over all values of Θ, using its distribution, there's no need to estimate the likelihood ratio, or to account for uncertainties, if a proper full Bayesian approach is chosen.

In the rest of the paper we only focus on frequentist methods to solve the rare haplotype problem, but a companion paper presents a similar study on Bayesian methods (Cereda, 2016a).

5.2.2 The frequentist perspective

The difference between frequentist and Bayesian methods regards parameters h and θ: for a frequentist they are fixed quantities, whose values correspond to, respectively, the unknown true value of θ and the correct hypothesis. One can see frequentist models as Bayesian models where the distributions chosen for Θ and H give probability one to values θ and h, respectively. Also, one can express the frequentist probability $\mathcal{Pr}$ in terms of the Bayesian probability Pr in the following way: $\mathcal{Pr}(\cdot \mid h) := \mathcal{Pr}_\theta(\cdot \mid h) = \Pr(\cdot | H=h, \Theta=\theta)$. For frequentist there is a true, 'physical' probability which governs the situation at hand: according to the prosecution this true probability is $\mathcal{Pr}_\theta(\cdot \mid h_p)$, while according to the defence it is $\mathcal{Pr}_\theta(\cdot \mid h_d)$, with θ set to its true (uknown) value.

Conditions **a** and **b** can be rephrased, in a frequentist language as:

Frequentist a. $\mathcal{Pr}_\theta(B=b \mid h_p) = \mathcal{Pr}_\theta(B=b \mid h_d)$, for all θ and b.

Frequentist b. $\mathcal{Pr}_\theta(E=e \mid B=b, h) = \mathcal{Pr}_\theta(E=e \mid h)$, for all θ, h, e, and b.

It holds that:

$$\mathcal{LR} = \frac{\mathcal{Pr}(D=d \mid h_p)}{\mathcal{Pr}(D=d \mid h_d)} = \frac{\mathcal{Pr}(E=e, B=b \mid h_p)}{\mathcal{Pr}(E=e, B=b \mid h_d)} = \frac{\mathcal{Pr}(E=e \mid B=b, h_p)}{\mathcal{Pr}(E=e \mid B=b, h_d)} \frac{\mathcal{Pr}(B=b \mid h_p)}{\mathcal{Pr}(B=b \mid h_d)}.$$

The index θ has been omitted for ease of notation. Thanks to conditions **Frequentist a** and **b**, the likelihood ratio can be expressed as

$$\mathcal{LR} = \frac{\mathcal{Pr}(E=e \mid h_p)}{\mathcal{Pr}(E=e \mid h_d)}. \tag{5.3}$$

Even though the two alternative ways of writing the likelihood ratio expressed by equations (5.2) and (5.3) are theoretically different, and mean two different things, they have the same value. This implies that part of the information, namely B, is not useful to discriminate between the two hypotheses of interest. Stated otherwise, when knowing θ, B is

irrelevant to determine the likelihood ratio, i.e. to decide about parameter h. However, it may play an important role in the estimation of parameter θ. For instance, getting back to the DNA example, the database (B) is often useful to estimate the frequencies of the different haplotypes.

Notice that, in order for (5.3) to hold, **b** can be modified to something less strong:

Frequentist b*. $\dfrac{\mathcal{P}r_\theta(E = e \mid B = b, h_p)}{\mathcal{P}r_\theta(E = e \mid B = b, h_d)} = \dfrac{\mathcal{P}r_\theta(E = e \mid h_p)}{\mathcal{P}r_\theta(E = e \mid h_d)}$ for all e, b, and θ.

which is equivalent to ask that updating the likelihood ratio for the observation of B to take into account the observation of E, does not change anything.

Furthermore, while conditions **a** and **b*** imply (5.3), the converse is not true. Formulation (5.3) is instead equivalent to a weaker condition, that is:

Frequentist c. $\mathcal{P}r_\theta(B = b \mid E = e, h_p) = \mathcal{P}r_\theta(B = b \mid E = e, h_d)$, for all θ.

This can be seen by the following alternative development of the likelihood ratio (θ omitted):

$$\mathcal{LR} = \frac{\mathcal{P}r(D = d \mid h_p)}{\mathcal{P}r(D = d \mid h_d)} = \frac{\mathcal{P}r(B = b \mid E = e, h_p)}{\mathcal{P}r(B = b \mid E = e, h_d)} \frac{\mathcal{P}r(E = e \mid h_p)}{\mathcal{P}r(E = e \mid h_d)} = \frac{\mathcal{P}r(E = e \mid h_p)}{\mathcal{P}r(E = e \mid h_d)}. \tag{5.4}$$

It follows that:

$$\mathbf{c} \Leftrightarrow \mathcal{LR} = \frac{\mathcal{P}r(E = e \mid h_p)}{\mathcal{P}r(E = e \mid h_d)}. \tag{5.5}$$

Notice that frequentists use a likelihood ratio $\mathcal{LR}_\theta$, which can be written in terms of the Bayesian LR as $\mathrm{LR}|\Theta = \theta$ (read "LR given θ"), and attempt to get close to θ by choosing some estimator $\widehat{\theta}$. This leads to the so-called *plug-in estimator* $\widehat{\mathcal{LR}_\theta} = \mathcal{LR}_{\widehat{\theta}} = \mathrm{LR}|(\Theta = \widehat{\theta})$. However, that's not the only option, as we will see for the method explained in Section 5.7.

It is important to notice that the frequentist approach may be represented by the same Bayesian network of Figure 5.1, where the states of nodes Θ and H are instantiated to particular values θ and h, respectively. This shows that actually the two approaches don't disagree on the structure of the model regarding E and B. Only, Bayesians add ingredients to the model by allowing Θ and H to have a distribution. Stated otherwise, the Bayesian approach is given by the very same frequentist conditions **a** and **b**, with the addition of condition **c** about the independence of Θ and H.

5.3 Data reduction

Let us denote with $\mathcal{D}$ all the data given to the expert in the form of a dossier, which he has to "translate" into a well-defined mathematical object. To evaluate the entirety of the data at the expert's disposal is often a delusion, from which the need of a reduction of $\mathcal{D}$ to something less informative, but of more feasible evaluation, which we denote as D. Often the database contains only information about a limited number of loci, and this implies that information about other loci of the crime stain can't be used. This constitutes already a first

reduction of the data. Other kinds of reductions are performed in order to gain in terms of precision of the estimates. Especially in a situation with many nuisance parameters, it can be wise to discard the part of data which primarily tells us about the nuisance parameters, and only indirectly about the ultimate question of interest (i.e., which hypothesis is more likely to be true). In fact, it could be very wise to reduce the data $\mathcal{D}$ to a much smaller amount of information, because the likelihood ratio based on the data reduction is much more precisely estimated than that based on all data. However, there's a limit to this: the reduction of $\mathcal{D}$ into D comes with a cost: the stronger the reduction, the less the corresponding likelihood ratio value is discriminating of the two hypotheses, because less information is less powerful to that purpose. We have to make a compromise between gain in precision and loss in strength of the evidence. This will be discussed more in detail in Section 5.8.

Once a particular reduction D has been defined, the frequentist likelihood ratio ($\mathcal{LR}$) can be defined as in (5.2). It is easy to understand that there isn't a unique way to reduce $\mathcal{D}$ and that each choice entails the definition of a different likelihood ratio. For instance, in the DNA example one can think of considering a profile made of more or less loci. Another kind of reduction will be presented in Section 5.7. Different choices of $D \subsetneq \mathcal{D}$ lead to different likelihood ratios. *Therefore it is better to refer to "a" likelihood ratio instead of to "the" likelihood ratio.* This was already stated in Dawid (2001), even though regarding hypotheses instead of data. In the literature different choices of $D \subsetneq \mathcal{D}$ and 'Pr' are proposed, each corresponding to a different likelihood ratio to be estimated. These choices are often only implicit and one of the aim of this research is to make explicit the reduction which corresponds to two selected methods, by looking for the corresponding E and B.

5.4 Different levels of uncertainty

The likelihood ratio measures the relative strength of support given by the data to an hypothesis over an alternative. Clearly, it is useful when there is uncertainty about which of the two hypotheses is true (to be more precise, it may also be the case that none of the alternatives is correct, and the likelihood ratio continues to be meaningful). Along with this first basic initial uncertainty about the state of the affairs, two more levels of uncertainty arise in the attempt of calculating the likelihood ratio.

For a frequentist statistician, the likelihood ratio is a ratio of probabilities based usually on a model $\mathcal{M}$ which is at best only a good approximation to the truth. Moreover, they have to estimate parameters of that model by fitting it to the data in some database. Stated otherwise, after a particular choice of what is the data D to be considered, a population model is to be chosen and its parameters estimated using a limited sample. Some forensic literature (Morrison, 2010; Stoel and Sjerps, 2012; Curran et al., 2002; Curran, 2005) already pointed out the necessity for uncertainty assessment in the likelihood ratio estimation, even though they don't differentiate among levels. On the other hand, for a true Bayesian statistician there's no need of estimation, and no additional levels of uncertainty to be added, since the definition of the Bayesian Pr already includes not only beliefs about chances when picking people from that population, but also beliefs about parameters of the models, and beliefs about models.

This discussion may hopefully put an end to the debate as to whether it makes sense to talk about 'estimation' and 'uncertainty assessment' for the likelihood ratio. Stoel and Sjerps (2012) believe that "there are strong arguments for the notion of a "true" but unknown value of the likelihood ratio, given the relevant hypotheses and background information, and that it is important to consider the uncertainty. Ignoring the uncertainty can be strongly misleading". This point of view is also shared in Sjerps et al. (2016). On the other hand, to talk about estimation of the likelihood ratio is defined as "internally inconsistent, and hence misconceived" by Taroni et al. (2016). Both the points of view are correct, if correctly put into context: if a frequentist approach is chosen it is sensible to talk about estimation and to deal with uncertainty assessment. On the other hand, in a full Bayesian context, they are misplaced.

Notice that Bayesianism is theoretically a very powerful interpretation of probability, but when it comes to apply Bayesian theory for practical purposes, even the most fervent Bayesian has to strike a balance between what is feasible and what is theoretically right and coherent according to the Bayesian perspective. He typically chooses a particular model as the correct one (as frequentists do), and/or he has to put convenient (rather than realistic) prior distributions on the parameters. Hence, whether Bayesian or frequentist approaches are chosen, the attempt to produce the likelihood ratio leads to several levels of uncertainty which should be accounted for.

We will now discuss the two additional levels of uncertainty mentioned before. The second level of uncertainty pertains to the choice of a particular population model, which is only an approximation of the truth. This level of uncertainty may be reduced using nonparametric methods, that are based on less assumptions.

Given a particular population model, the third level of uncertainty pertains to the fact that the population parameters are not known. This may involve estimation of parameters (such as in the Discrete Laplace method of Section 5.6) or the direct estimation of the probabilities of interest (as in the Generalized Good method described in Section 5.7) and the quality of the estimates severely depends on the size of the available databases. This level of uncertainty pertains both to parametric and nonparametric methods.

The evidential value reported depends on all the levels of uncertainty which afflict the estimation of the likelihood ratio. Thus, it is of the utmost importance to report the likelihood ratio value along with (1) an explicit definition of which data D we want to evaluate through that likelihood ratio, and (2) a discussion (and if possible quantification) of the levels of uncertainty that afflict the reported value.

5.4.1 Estimating the weight of evidence

Instead of estimating the likelihood ratio, it is more sensible to directly estimate its logarithm, sometimes called *relevance ratio* or *weight of evidence* (Good, 1950; Aitken et al., 1998; Aitken and Taroni, 2004). This is because the interpretation of the likelihood ratio values goes through orders of magnitude 10, and when a value is reported, it is important to control the relative error, rather than the absolute error. In fact, the first is meaningful in itself while the second depends on the particular value of the likelihood ratio. For the very same reasons why

the verbal equivalent scale (Aitken et al., 1998) is based on logarithm. Furthermore, both the odds form of Bayes' theorem and the formula to combine likelihood ratios from independent pieces of evidence involve a multiplicative relationship that becomes a more handy additive relation if logarithm is taken (Schum, 1994). Moreover, the logarithm helps in presenting large numbers in a compact way, of more easy comprehension, and it is symmetric with respect to prosecution's and defence's hypothesis: this may be useful if one wants to invert the weight of evidence to consider the defence's proposition (Aitken and Taroni, 2004).

5.5 The rare Y-STR haplotype problem

Consider the situation in which a piece of evidence is recovered at the crime scene, and a suspect turns out to have the same analysed characteristics (for instance the same DNA profile) as the crime scene evidence. Prosecution claims that the suspect left the evidence, defence claims that someone else (with the same DNA profile) left it. The strength of the match to discriminate between the competing hypotheses is evaluated by comparing how probable this is under each of the hypotheses. This depends on the proportion of individuals in possess of the same profile in the population of possible perpetrators: the rarer the profile the more the suspect is in trouble. This proportion is usually unknown, the only available data being a sample of DNA profiles from the population, in the form of a reference database. The *naive estimator* uses the relative frequency of the profile in the database as estimate for θ. Problems arise when this frequency is 0, the so-called "rare type match". This problem is so substantial that it has been defined "the fundamental problem of forensic mathematics" by Brenner (2010). As an alternative to the empirical frequency estimator, one can use the *add-constant* estimators, which adds a constant to the count of each type, included the unseen ones. The most well known is the *add-one* estimator, due to Laplace (1814), and the *add-half* estimator of Krichevsky and Trofimov (1981). However, to use these methods one needs to know the number of possible unseen types and there are problems if this number is large compared to the sample size (see Gale and Church (1994) for additional discussion). Another possibility is the 'rule of three', proposed by Louis (1981). It states that $3/n$ is a good approximation of the 95% upper bound for the frequency, if n is the size of the database.

Of interest for this paper is the nonparametric *Good Turing estimator* of Good (1953), based on an intuition on A. M. Turing. It is an estimator for the total unobserved probability mass which is based on the proportion of singletons in the sample. For a comparison between *add one* and *Good-Turing* estimator, see Orlitsky et al. (2003).

The *naive estimator* and the *Good Turing estimator* are in some sense complementary (Anevski et al., 2013): the first gives a good estimate for the observed types, and the second for the probability mass of the unobserved ones. Lastly, the *high profile estimator*, introduced by Orlitsky et al. (2004), extends the tail of the *naive estimator* to the region of unobserved types. This estimator has been improved by Anevski et al. (2013) that also provides the consistency proof.

The rare type match problem is common, for instance, in case a new kind of forensic evidence is involved, and for which the available database size is still limited. One example is the case

of DIP-STR markers (e.g. Cereda et al., 2014a). The same happens when Y-chromosome (or mitochondrial) DNA profiles are used: because of the lack of recombination involved when offspring DNA is generated from the DNA of the parents, the haplotype must be treated as a unit (the match probability can't be obtained by multiplication across loci) and the set of possible haplotypes is extremely large. As a consequence, most of the Y-STR haplotypes are not represented in the database.

In the rest of the paper, the Y-STR profile example will be retained as an extreme but common and important way in which the problem of assessing the evidential value of rare type match can arise. Literature provides some examples of approaches to evaluate it for the rare Y-STR haplotypes match: Egeland and Salas (2008), the κ method Brenner (2010, 2014), the coalescent theory method (Andersen et al., 2013a), the haplotype surveying method (Roewer et al., 2000; Krawczak, 2001; Willuweit et al., 2011), and the Discrete Laplace method (Andersen et al., 2013b) (not directly proposed for the rare haplotype case but usable for that purpose). As already mentioned, Cereda (2016a) discusses the full Bayesian approach to this problem.

Bayesian nonparametric estimators for the probability of observing a new type have been proposed by Tiwari and Tripathi (e.g. 1989); Lijoi et al. (e.g. 2007); Favaro et al. (e.g. 2009). However, for the likelihood ratio assessment it is required not only the probability of observing a new species but also the probability of observing this same species twice (according to the defense the crime stain profile and the suspect profile are two independent observations). Cereda (2016c) is the first paper that addresses the problem of likelihood ratio assessment in the rare haplotype case using Bayesian nonparametric models.

The present paper analyses two frequentist methods, the Discrete Laplace method, and a generalization of the Good Turing, making explicit the corresponding definitions of D, E and B, and providing a study on the different levels of uncertainty arising for each.

5.6 The Discrete Laplace Method

A discrete random variable X is said to follow the Discrete Laplace distribution $\mathrm{DL}(p, y)$, with dispersion parameter $p \in (0, 1)$, and location parameter $y \in \mathbb{Z}$, if its probability density is defined as

$$f(x \mid p, y) = \left(\frac{1-p}{1+p}\right) p^{|x-y|}, \quad \forall x \in \mathbb{Z}.$$

This is used in Andersen et al. (2013b) to model the distribution of single locus Y-STR haplotype in some subpopulation, which is thus assumed to be distributed around a modal allele (represented by the location parameter y).

Each haplotype is actually composed by r loci. Let denote with $\mathbf{X} = (X_1, X_2, ..., X_r)$ the random variable which describes an r-loci haplotype configuration. Moreover, there may be c different subpopulations to take into consideration. By making the strong assumption of independence between loci, within the same subpopulation, the following density is used to describe the probability that $\mathbf{X} = \mathbf{x}$:

$$f(\mathbf{x} \mid \{\mathbf{y_j}\}_j, \{\mathbf{p_j}\}_j) = \sum_{j=1}^{c} \tau_j \prod_{k=1}^{r} f(x_k \mid y_{jk}, p_{jk}),$$

where, for each j, τ_j is the probability a priori of generating from the jth subpopulation, while $\mathbf{p_j} = (p_{j1}, p_{j2}, ..., p_{jr})$ and $\mathbf{y_j} = (y_{j1}, y_{j2}, ..., y_{jr})$ represent the dispersion and location parameters, respectively, of the jth subpopulation. Andersen et al. (2013b) propose to estimate all these parameters by using using the EM algorithm (Dempster et al., 1977). The initial subpopulation centres are chosen by PAM algorithm (Kaufman and Rousseeuw, 2009) and the number of them by the Bayesian Information Criteria (BIC) (Schwarz, 1978).

5.6.1 The choice of D in the Discrete Laplace Method

The choice of D which underlies the Discrete Laplace method, when used to address the rare haplotype match problem is:

- D_{DL} = The particular haplotype x of the suspect and of the stain, along with a database which is a sample from the population of possible perpetrators.

This method allows to evaluate the data in the light of the usual hypotheses of interest in the DNA example (see Section 5.2). D_{DL} can be split into E_{DL} and B_{DL}, in the following way:

- E_{DL} = the particular haplotype x of the stain (E_t) and of the suspect (E_s).
- B_{DL} = reference sample from the population of possible perpetrators (i.e. database).

The vector containing the frequencies of all haplotypes in the population of reference can be thought of as the nuisance parameter θ of this model. Conditions **a.** and **b.** presented in Section 5.2 are valid for E_{DL}, B_{DL}, θ, and h, thus the following likelihood ratio (where θ is again omitted) corresponds to this choice of data, evidence, background and model:

$$\begin{aligned} \mathcal{LR}_{\text{DL}} &= \frac{\mathcal{P}r(D_{\text{DL}} = d \mid h_p)}{\mathcal{P}r(D_{\text{DL}} = d \mid h_d)} = \frac{\mathcal{P}r(E_t = x \mid E_s = x, h_p)\mathcal{P}r(E_s = x \mid h_p)}{\mathcal{P}r(E_t = x \mid E_s = x, h_d)\mathcal{P}r(E_s = x \mid h_d)} \\ &= \frac{\mathcal{P}r(E_t = x \mid E_s = x, h_p)}{\mathcal{P}r(E_t = x \mid h_d)} = \frac{1}{f_x}. \end{aligned} \tag{5.6}$$

Here, f_x is the frequency of the haplotype x in the population of reference. The second equality is due to conditions **a** and **b** discussed in Section 5.2.2, while the forth one is justified by the fact that the distribution of the haplotype of the suspect does not depend on which hypothesis is correct, and that, when θ is fixed (as in the frequentist approach which we are considering) and under h_d, E_t is independent of E_s. The weight of evidence is thus

$$\log_{10} \mathcal{LR}_{\text{DL}} = \log_{10} \frac{1}{f_x}. \tag{5.7}$$

The frequency f_x can be estimated by $\widehat{f_x}$, using the Discrete Laplace method. This brings to the following plug-in estimator for $\log_{10} \mathcal{LR}_{\text{DL}}$:

$$\widehat{\log_{10} \mathcal{LR}_{\text{DL}}} = \log_{10} \frac{1}{\widehat{f_x}}.$$

Notice that the Discrete Laplace method uses the database to estimate the number of subpopulations and all the parameters in the model, and this is where B_{DL} comes into play again.

5.6.2 Quantifying the uncertainty of the Discrete Laplace method

We quantify the uncertainty of this method comparing the distribution of $\widehat{\log_{10} \mathcal{LR}_{\text{DL}}} = \log_{10} \frac{1}{\widehat{f_x}}$ with the distribution of the "true" $\log_{10} \mathcal{LR}_{\text{DL}} = \log_{10} \frac{1}{f_x}$. f_x is not known, but we have a database of approximately 19,000 Y-STR 23-loci profiles from 129 different locations in 51 countries in Europe (Purps et al., 2014)[1], which we can pretend contains the whole population of interest for our case. We will consider only 7 loci out of 23 and perform the following experiment: we sample a small database of size $N = 100$, along with a new haplotype (not observed in the small database), and calculate the estimate $\log_{10} \frac{1}{\widehat{f_x}}$. Then, we can use the relative frequency of the haplotype x in the big database as the true one, f_x to obtain $\log_{10} \frac{1}{f_x}$

This process can be repeated many times (for instance $M = 1000$ samplings of small databases of size $N = 100$ and, for each, a never observed haplotype).

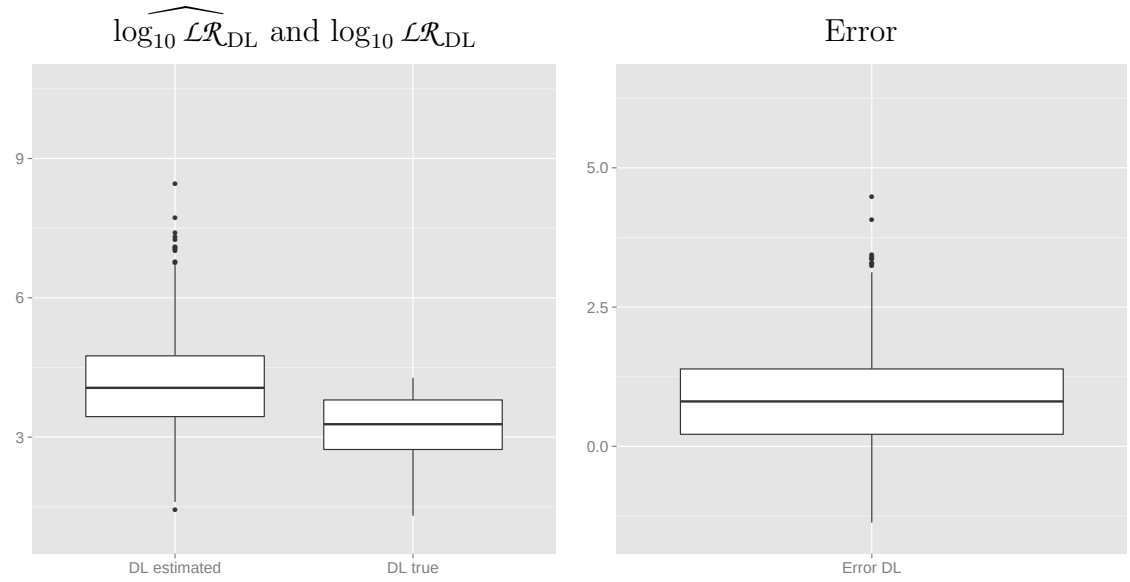
In estimating $\log_{10} \mathcal{LR}_{\text{DL}}$ via $\widehat{f_x}$, one has the choice between adding the haplotype x to the small database before estimating parameters of the Discrete Laplace distribution, or not. In a full Bayesian approach the right thing to do is to add the profile to the database. This is shown in Cereda (2016a), and we believe that it is the good thing to do also in a frequentist framework. In fact, experiments show that to add or not the haplotype to the database does not make much difference. Table 5.1 and Figure 5.2 (left part) compare the distributions of $\log_{10} \mathcal{LR}$ and $\widehat{\log_{10} \mathcal{LR}_{\text{DL}}}$, using 7 loci. The same experiment has been carried out for 10 and 3 loci, but not reported in details.

	Min	1st Qu.	Median	Mean	3rd Qu.	Max	s.d.
$\log_{10} \mathcal{LR}_{\text{DL}}$	1.305	2.733	3.277	3.272	3.800	4.277	0.666
$\widehat{\log_{10} \mathcal{LR}_{\text{DL}}}$	1.432	3.441	4.061	4.114	4.750	8.452	1.017
Error e_{DL}	-1.37	0.217	0.807	0.842	1.39	4.476	0.863

Table 5.1: Summaries of the distribution of $\log_{10} \mathcal{LR}_{\text{DL}}$, $\widehat{\log_{10} \mathcal{LR}_{\text{DL}}}$, and of the error e_{DL}.

The error of the Discrete Laplace method can be defined as $e_{\text{DL}} := \widehat{\log_{10} \mathcal{LR}_{\text{DL}}} - \log_{10} \mathcal{LR}_{\text{DL}}$. It measures how much the estimated distribution differs from the true one. Table 5.1 and Figure 5.2 (right part) show the distribution of the error. One can see that it can attain up to about 4 orders of magnitude. The distribution of the error is mostly located on positive values, which means that, more often than not, $\widehat{\log_{10} \mathcal{LR}_{\text{DL}}}$ overestimates $\log_{10} \mathcal{LR}_{\text{DL}}$. The

[1] A clean version of the database is provided by Mikkel Meyer Andersen (`http://people.math.aau.dk/~mikl/?p=y23`).



(a) (b)

Figure 5.2: Discrete Laplace method. Boxplots comparing the distributions of $\log_{10} \mathcal{LR}_{\mathrm{DL}}$ and $\widehat{\log_{10} \mathcal{LR}_{\mathrm{DL}}}$ (left) and the error $e_{\mathrm{DL}} = \widehat{\log_{10} \mathcal{LR}_{\mathrm{DL}}} - \log_{10} \mathcal{LR}_{\mathrm{DL}}$ (2nd column).

standard deviation of the error is small, thereby e_{DL} does not move too much away from its mean, which is about 0.842.

Motivated by the discussion of Section 5.4, we now analyze the different levels of uncertainty which affect the error. The second level of uncertainty is introduced when the Discrete Laplace model, along with all its set of assumptions, is chosen to model the distribution of single locus haplotypes, which in reality do not follow a Discrete Laplace distribution.

The third level of uncertainty pertains to the estimation of the parameters of the model (c, p, y, τ). Here, the databases used to estimate the parameters of the Discrete Laplace model are probably too small ($N = 100$) with regard to 7 loci.

To decrease both sources of error, one can reduce the number of analyzed loci to 3. The population becomes less sparse, and the databases big enough. We performed this experiment and indeed the error decreased a great deal. However, the basic level of uncertainty (see Section 5.4) is increased inasmuch the data becomes less effective to discern between the two hypotheses. On the other hand, the same experiment with 10 loci lead to obtain more powerful likelihood ratios, but less precise.

The second level of uncertainty can be made harmless assuming an infinite number of subpopulations, since in this way the model will perfectly fit any population, even though, with this solution, the number of parameters will increase, along with the third level of uncertainty.

It is worth underlining that the results of our simulations do not mean that the Discrete Laplace method is wrong on the whole, but they show that a blind use of this method is dangerous. We are applying this method to the specific case of the rare haplotype match,

using s of size 100, and a rather sparse population: maybe this method was never intended to be used for such small databases, and maybe it can be modified in more clever ways to that purpose.

5.7 The Generalized Good method

Based on Good (1953), we now propose a nonparametric estimator for the weight of evidence. This is a very good example of data reduction, since $\mathcal{D}$ is here reduced to a greater extent than it was done for the Discrete Laplace method. Indeed, the specific haplotype x of the crime stain and of the suspect is ignored, retaining only the fact that they match and the fact that this profile has not been observed yet in the database.

Stated otherwise,

- D_{GG} = the haplotype of the suspect matches the haplotype of the crime stain and it is not in the database.

Consider the following mathematical description: the database of size N can be seen as an i.i.d. sample $(Y_1, Y_2, ..., Y_N)$ from species $\{1, 2, ..., S\}$, with probabilities $(p_1, p_2, ...p_S)$. Hence, the suspect's profile can be though of as the $N+1$st i.i.d. observation. The crime stain's profile is the $N+2$nd observation. According to the defence it is again an i.i.d. draw from $(p_1, p_2, ...p_S)$, while according to prosecution it is, with probability one, equal to the value of Y_{N+1}.

To make the notation less cumbersome we are using

$$\begin{aligned}
&\mathcal{Y}_N := (Y_1, Y_2, ..., Y_N),\\
&\mathcal{Y}_{i,N} := (Y_1, Y_2, ..., Y_{i-1}, Y_{i+1}, ..., Y_N),\\
&\mathcal{Y}_{(i,j),N} := (Y_1, Y_2, ..., Y_{i-1}, Y_{i+1}, ..., Y_{j-1}, Y_{j+1}, ..., Y_N), \quad \forall i < j.
\end{aligned}$$

Moreover, for any random variable Y, and any couple of sets A and B, $\mathbf{1}_{A\cap B^c}(Y)$ is a random variable which has value 1 if Y belongs to the set A and not to the set B, and zero otherwise.

The likelihood ratio for this reduction of the data can be thus written as

$$\mathcal{LR}_{\text{GG}} = \frac{\mathcal{P}r(Y_{N+1} \notin \mathcal{Y}_N, Y_{N+1} = Y_{N+2} \mid h_p)}{\mathcal{P}r(Y_{N+1} \notin \mathcal{Y}_N, Y_{N+1} = Y_{N+2} \mid h_d)} = \frac{\mathcal{P}r(Y_{N+1} \notin \mathcal{Y}_N \mid h_p)}{\mathcal{P}r(Y_{N+1} \notin \mathcal{Y}_N, Y_{N+1} = Y_{N+2} \mid h_d)}.$$

From now on, we are presenting results regarding a general database size $N > 2$, and general random variables $Y_1, ..., Y_N$, i.i.d. from $(p_1, p_2, ..., p_S)$. The following notation is used:

$$\begin{aligned}
&\theta_1(N; p_1, p_2, ..., p_S) := \mathcal{P}r(Y_N \notin \{Y_1, Y_2, ..., Y_{N-1}\}),\\
&\theta_2(N; p_1, p_2, ..., p_S) := \mathcal{P}r(Y_N \notin \{Y_1, Y_2, ..., Y_{N-2}\}, Y_N = Y_{N-1}).
\end{aligned}$$

Theorem 1. *An unbiased estimator for $\theta_1(N; p_1, p_2, ..., p_S)$ is $\widehat{\theta_1}(N) = N_1/N$, where N_1 is the number of singletons in the database.*

Proof.

$$\theta_1(N; p_1, p_2, ..., p_S) = \mathcal{P}r(Y_N \notin \mathcal{Y}_{N-1}) = \mathbb{E}(\mathbf{1}_{(\mathcal{Y}_{N-1})^c}(Y_N)) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}(\mathbf{1}_{(\mathcal{Y}_{i,N})^c}(Y_i))$$
$$= \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}_{(\mathcal{Y}_{i,N})^c}(Y_i)\right) = \mathbb{E}\left(\frac{N_1}{N}\right),$$

where the last equality is due to the fact that the function $\mathbf{1}_{(\mathcal{Y}_{i,N})^c}(Y_i)$ has value 1 for every singleton of the database: the sum is thus the number of singletons (N_1). □

Theorem 2. *An unbiased estimator for $\theta_2(N; p_1, p_2, ..., p_S)$ is $\widehat{\theta}_2(N) = 2N_2/N(N-1)$, where N_2 is the number of doubletons in the database.*

Proof.

$$\theta_2(N; p_1, p_2, ..., p_S) = \mathcal{P}r(Y_N \notin \{\mathcal{Y}_{N-2}\}, Y_N = Y_{N-1}) = \mathbb{E}(\mathbf{1}_{\{Y_{N-1}\cap(\mathcal{Y}_{N-2})^c\}}(Y_N))$$
$$= \frac{2}{N(N-1)}\sum_{i<j}\mathbb{E}(\mathbf{1}_{\{Y_j\cap(\mathcal{Y}_{(i,j),N})^c\}}(Y_i)) = \mathbb{E}\left(\frac{2}{N(N-1)}\sum_{i<j}\mathbf{1}_{\{Y_j\cap(\mathcal{Y}_{(i,j),N})^c\}}(Y_i)\right)$$
$$= \mathbb{E}\left(\frac{2N_2}{N(N-1)}\right),$$

where the last equality is due to the fact that the function $\mathbf{1}_{\{Y_j\cap(\mathcal{Y}_{(i,j),N})^c\}}(Y_i)$ has value 1 for each of the N_2 doubletons of the database. □

The two previous theorems can be easily generalized to θ_m defined as $\theta_m(N; p_1, p_2, ..., p_S) := \mathcal{P}r(Y_N \notin \mathcal{Y}_{N-m}, Y_N = Y_{N-1} = .. = Y_{N-m+1})$.

Now we can estimate $\log_{10}\mathcal{LR}_{\text{GG}}$ in the following way:

$$\log_{10}\mathcal{LR}_{\text{GG}} = \log_{10}\frac{\mathcal{P}r(Y_{N+1} \notin \mathcal{Y}_N \mid h_p)}{\mathcal{P}r(Y_{N+1} \notin \mathcal{Y}_N, Y_{N+1} = Y_{N+2} \mid h_d)} \approx \log_{10}\frac{\mathcal{P}r(Y_N \notin \mathcal{Y}_{N-1})}{\mathcal{P}r(Y_N \notin \mathcal{Y}_{N-2}, Y_N = Y_{N-1})}$$
$$\approx \log_{10}\frac{\theta_1(N; p_1, p_2, ..., p_S)}{\theta_2(N; p_1, p_2, ..., p_S)}.$$

Thus, we propose the following estimator for the weight of evidence:

$$\widehat{\log_{10}\mathcal{LR}}_{\text{GG}} = \log_{10}\frac{\widehat{\theta}_1(N)}{\widehat{\theta}_2(N)} = \log_{10}\frac{(N-1)N_1}{2N_2} \approx \log_{10}\frac{NN_1}{2N_2}. \tag{5.8}$$

Notice that there are two kinds of approximation steps: a mathematical approximation of $\theta_1(N+1; p_1, p_2, ..., p_S)$ with $\theta_1(N; p_1, p_2, ..., p_S)$, which should hardly make any difference, for reasonably large N, and a statistical estimation of $\theta_1(N; p_1, p_2, ..., p_S)$ using an unbiased estimator (and similarly for θ_2).

It is important to underline that, due to Jensen's inequality, the estimators $\log_{10}\widehat{\theta}_1$ and $\log_{10}\widehat{\theta}_2$ are not unbiased for $\log_{10}\theta_1$ and $\log_{10}\theta_2$, but it will be shown by simulations that

$\widehat{\log_{10} \mathcal{LR}_{\text{GG}}}$ is approximately unbiased for $\log_{10} \mathcal{LR}_{\text{GG}}$. However, the point is not to find an unbiased estimator, but one with a small error rate.

Notice that in order to estimate $\log_{10} \mathcal{LR}_{\text{GG}}$ it is not necessary to use all the information contained in the database, but only N, N_1, and N_2, that is the number of singletons and doubletons in the database. The nuisance parameter of the model is the vector θ containing the frequencies of the Y-STR haplotypes in the population of interest. θ_1 and θ_2 are functions of θ.

The limitation of this method is that it cannot be used if $N_2 = 0$ (this corresponds to an infinite likelihood ratio) and it does not perform well also in case the number of singletons is very small or zero. We believe it can be improved and extended by smoothing techniques (Good, 1953; Anevski et al., 2013), but we are going to ignore this problem.

The 'κ-method' of Brenner (Brenner, 2010) is based on an analogous line of reasoning. It estimates the likelihood ratio as $\widehat{\mathcal{LR}}_{\kappa} = \frac{N^2}{N-N_1}$. However, in the derivation of this estimator, there is an approximation involved, based on assumptions which are not always satisfied, leading sometimes to anti-conservatism (see also the discussion in Buckleton et al. (2011), and the answer in Brenner (2014)). In particular, Brenner (2014) provides a pathological population where the approximation does not hold, while showing empirical evidence that for Fisher-Wright populations the condition is fulfilled. Our method is, on the other hand, based on a principled derivation of the estimator of equation (5.8), which is similar to Brenner's one under the following conditions: there are almost only singletons and doubletons in the database, and $N_1 \gg N_2$.

These assumptions are typically satisfied, explaining why Brenner's method often works. They also constitute a good description of when it does not work.

Lastly, we remark that this method can be generalized in the obvious way, to the case in which the haplotype is indeed in the database. Moreover, this method is suitable to be directly applied to different kinds of evidence.

5.7.1 Quantifying the uncertainty of the GG method

As we did in Section 5.6.2, we want to quantify the uncertainty of this method. One way is to compare the distribution of

$$\widehat{\log_{10} \mathcal{LR}_{\text{GG}}} = \log_{10} \frac{NN_1}{2N_2},$$

with the distribution of the "true"

$$\log_{10} \mathcal{LR}_{\text{GG}} = \log_{10} \frac{\mathcal{P}r(Y_{N+1} \notin \mathcal{Y}_N)}{\mathcal{P}r(Y_{N+1} \notin \mathcal{Y}_N \cap Y_{N+1} = Y_{N+2})} := \log_{10} \frac{\theta_1}{\theta_2}.$$

Actually, the latter is not a distribution, but a single value, unknown. Again, we pretend that the database of Purps et al. (2014) contains the profiles of the whole population, to find out the 'true' θ_1 and θ_2, restricting our simulations to 7 loci. To do so, we sample M small

databases of size $N = 100$, along with two other haplotypes. θ_1 is the proportion of times in which the $(N+1)$st haplotype is a new one (i.e., not one of the previous N), and θ_2 is the proportion of times in which the $(N+2)$nd is equal to the $(N+1)$st, and different from the first N observations. From our simulations, we used $M = 100,000$, and we obtained θ_1, θ_2, and $\log_{10} \mathcal{LR}$ as in Table 5.2.

θ_1	θ_2	True $\log_{10} \mathcal{LR}_{\text{GG}}$
0.748	0.0012	2.78

Table 5.2: Values of θ_1 and θ_2 and of $\log_{10} \mathcal{LR}_{\text{GG}}$ obtained by simulation, assuming that the database of Purps et al. (2014) contains the whole population of interest.

The distribution of $\widehat{\log_{10} \mathcal{LR}}_{\text{GG}} = \log_{10} \frac{NN_1}{2N_2}$ can be obtained by sampling $M = 100,000$ databases of size $N = 100$. Out of 100,000 databases, 121 had $N_2 = 0$. These have been removed from the data, and we acknowledge that this creates unfairness to the Discrete Laplace method, but we believe that this occurs frequenty enough not to affect very strongly the comparison. Figure 5.3 shows the distribution of the estimator $\widehat{\log_{10} \mathcal{LR}}_{\text{GG}}$ around the true

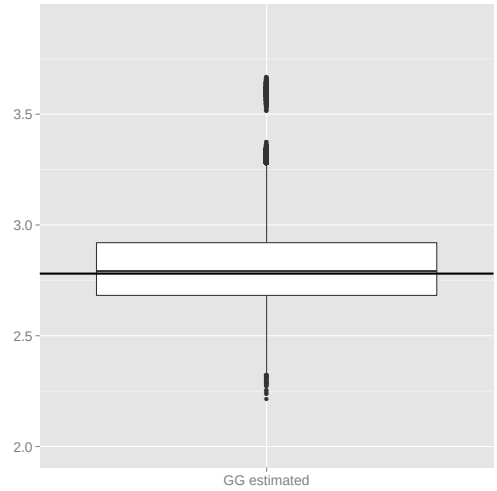


Figure 5.3: Boxplots of the distributionf of $\widehat{\log_{10} \mathcal{LR}}_{\text{GG}}$ around the true value $\log_{10} \mathcal{LR}_{\text{GG}}$ (black line).

value (black line). The error of the Generalized Good method, defined as $e_{\text{GG}} = \widehat{\log_{10} \mathcal{LR}}_{\text{GG}} - \log_{10} \mathcal{LR}_{\text{GG}}$, tells us how much the estimator differs from the true value.

	Min	1st Qu.	Median	Mean	3rd Qu.	Max	sd
$\log_{10} \mathcal{LR}_{\text{GG}}$	2.78	2.78	2.78	2.78	2.78	2.78	0
$\widehat{\log_{10} \mathcal{LR}}_{\text{GG}}$	2.215	2.682	2.792	2.818	2.920	3.668	0.198
Error e_{GG}	-0.566	-0.098	0.0112	0.038	0.14	0.887	0.198

Table 5.3: Summaries of the distribution of $\widehat{\log_{10} \mathcal{LR}}_{\text{GG}}$, of $\log_{10} \mathcal{LR}_{\text{GG}}$, and of the error e_{GG}.

Table 5.3 provides the summaries for $\widehat{\log_{10} \mathcal{LR}_{\mathrm{GG}}}$, and for the error e_{DL}. We don't provide the plots for the distribution of e_{GG} since they are identical to those in Figure 5.3, shifted of $\log_{10} \mathcal{LR}_{\mathrm{GG}}$.

One can see that the error can attain up to about 0.9 orders of magnitude. The distribution of the error is mostly located on the positive values, which means that, more often than not, $\widehat{\log_{10} \mathcal{LR}_{\mathrm{GG}}}$ overestimates $\log_{10} \mathcal{LR}_{\mathrm{GG}}$. The standard deviation of the error is small, thereby e_{GG} does not move too much away from the mean, which is about 0.038. If compared to the error of the Discrete Laplace method, one can conclude that here we get a better estimator in terms of accuracy, since the error ranges over more restrained values and the standard deviation is much smaller. However, it is important to keep in mind that they are not different estimators of the same quantity, but different estimators of different quantities, since the reduction of data used by the Generalized Good method, which allows to obtain accuracy in the estimates is less strong to discern between the two hypotheses.

5.8 Choosing and comparing methods

In comparing the two methods one can consider the precision with respect to what the method is trying to estimate, quantified by the errors e_{DL}, and e_{GG}. These errors are due to the two second and third level of uncertainty described in Section 5.4, and decrease sensibly if data is reduced. This is why, under this aspect, the Generalized Good is to be preferred to the Discrete Laplace, and for the latter a fewer number of loci is to be preferred. However, it is not correct to believe that the greater the reduction, the better is the method. To reduce means to lose information, and thus to diminish the capability of the method to distinguish between the hypotheses at stake (the first, or basic level of uncertainty). In order to investigate this loss, one can compare each method to the likelihood ratio $1/f$ (where f is the population frequency of the matching haplotype), which can be considered the hallmark in a population with no substructure. Comparing Table 5.1 with Table 5.2 one can see for instance that choosing the Generalized Good one loses on average around 0.5 (in logarithmic scale) in terms of strenght of data to discriminate between hypotheses. This is a small disadvantage for the prosecution, while everybody gain in terms of precision with respect to the true $\log_{10} \mathcal{LR}_{\mathrm{GG}}$. As a last remark, we invite the reader to realize that the Discrete Laplace method is better inasmuch it can always be used. On the other hand, for the Generalized Good, we had to remove 121 experiments where $N_2 = 0$.

5.9 Remark and conclusion

The aim of this paper could, at first sight, be considered that of offering two additional frequentist methods to address the issue of the likelihood ratio calculation in case of rare type match. However, a careful reader may have realized that these methods also constitute two interesting examples to apply the guidelines exposed in the opening sections. In particular, two important facts are pointed out in Sections 5.3 and 5.4: first, it is more sensible to talk about "a" likelihood ratio instead of "the" likelihood ratio, and second, a quantification of

the error involved in the estimation is to be provided along with the estimated likelihood ratio value.

Moreover, it is explained that sometimes it is possible to the break down data to be evaluated, into E (which is sufficient for H), and B (which is irrelevant for H). The Discrete Laplace method (developed in Section 5.6) is a good example where this distinction can be done, while the same is not true for the Generalized Good method (Section 5.7).

Lastly, this paper wants to get across the message that reducing the data to a smaller extent is sometimes not only necessary, but also desirable in terms of exactitude of the estimates, as proved by the comparison between the Discrete Laplace method (less reduction, less precision of the estimates) and the Generalized Good method (stronger reduction, more precision of the estimates). In this respect we disagree with Buckleton et al. (2011) who, talking about Brenner's method, state that 'there is a merit focussing in the type or name of a lineage marker". Although we agree that "such ignorance of type implies a substantial loss of information", it may allow a large gain in precision.

The take home message is that to choose the best method is clearly a very delicate task. One has to consider many different aspects, and look for a compromise which is acceptable for the specific application at hand. It is important to realise that in this paper we study a very extreme situation with very small databases and a possibly unrealistic population, for which the Generalized Good seemed to be the best compromise. Clearly, there are no possible general conclusions to be given, if not that at each new situation one has to reconsider all these aspects, and weigh them.

Acknowledgements

The generalized Good method described here was suggested by Richard Gill and presented in several conference lectures, [2]. I am indebted to Charles Brenner (the first to use the 'fundamental problem' name), and to Ronald Meester, for the useful discussions about this paper, which lead to many improvements. This research was supported by the Swiss National Science Foundation, through grants no. 105311-1445570 and 10531A-156146/1, and carried out in the context of a joint research project, supervised by Franco Taroni (University of Lausanne, Ecole des sciences criminelles, Faculté de droit, des sciences criminelles et d'administration publique), and Richard Gill (Mathematical Institute, Leiden University).

[2] see for instance
http://www.slideshare.net/gill1109/the-fundamental-problem-of-forensic-statistics-38322519

Chapter 6

Bayesian approach to LR for the rare match problem

This chapter is based on:
Cereda, G. (2016) Bayesian approach to LR assessment in case of rare type match: careful derivation and limits. *Statistica Neerlandica*, In Press.

Abstract

The likelihood ratio (LR) is largely used to evaluate the relative weight of forensic data regarding two hypotheses and for its assessment Bayesian methods are widespread in the forensic field. However, the Bayesian 'recipe' for the LR presented in most of the literature consists of plugging-in Bayesian estimates of the involved nuisance parameters into a frequentist-defined LR: frequentist and Bayesian methods are thus mixed, giving rise to solutions obtained by hybrid reasoning. This paper provides the derivation of a proper Bayesian approach to assess LRs for the 'rare type match problem', the situation in which the expert wants to evaluate a match between the DNA profile of a suspect and that of a trace from the crime scene, and this profile has never been observed before in the database of reference. LR assessment using the two most popular Bayesian models (beta-binomial and Dirichlet-multinomial) is discussed and compared to corresponding plug-in versions.

6.1 Introduction

One of the main challenges of forensic science is that of properly evaluating the match between the characteristics of a crime stain (for instance a Y-STR profile) and the corresponding characteristics of some material from a known source (for instance from a suspect). Typically, a couple of mutually exclusive hypotheses is defined, of the kind of 'the crime stain came from the suspect' (h_p) and 'the crime stain came from an unknown donor' (h_d). The forensic expert is given some data D which can typically be split into *evidence*, data directly related

to the crime, and *background*, additional data not directly related to the crime and only pertaining some nuisance parameter θ involved in the assessment of the likelihood ratio. Evidence and background data will be modelled in this paper through random variables E and B respectively. In particular, we are interested in the situation in which the forensic expert is asked to evaluate the match between the Y-STR profile of a suspect and the Y-STR profile of a stain found at the crime scene. It is intuitive to understand that (one of) the nuisance parameter(s) involved in this evaluation is the proportion of people with the same profile in the relevant population: the more this profile is rare the more the suspect is in trouble. This proportion is unknown and thus the expert is given (or asks for) a database containing a list of Y-STR profiles from a sample from the relevant population. The main difference between the frequentist and the Bayesian methodology is that the first considers the nuisance parameter θ and the correct hypothesis h as fixed (without distribution) unknown quantities, while the second models the expert's uncertainty about the value of θ and h through random variables Θ and H, whose prior distributions reflect prior beliefs of the expert. The largely accepted method for evaluating the data in order to discriminate between the two hypotheses of interest, is the calculation of the *Bayes factor* (BF), regularly called *likelihood ratio* (LR) in forensic context and defined as the ratio of the probabilities of observing the data under the two competing hypotheses:

$$\mathrm{LR} = \frac{\Pr(E = e, B = b \mid H = h_p)}{\Pr(E = e, B = b \mid H = h_d)} = \frac{\Pr(E = e \mid B = b, H = h_p)}{\Pr(E = e \mid B = b, H = h_d)}. \tag{6.1}$$

In the Bayesian framework (the one of interest for this paper) Pr is the joint distribution of all the random variables in the model (E, B, H, and Θ). The second equality is justified by the fact that background data, by definition, is independent of H. The formula on the right looks more similar to the one usually presented in the literature of reference (e.g., Aitken and Taroni (2004), Taroni et al. (2014)), where only E is evaluated, and B is either referred to as I or not explicitly written. In the latter case, Pr has to be thought of as representing any uncertainty conditional on $B = b$.

On the other hand, frequentists, who consider θ and h as fixed quantities, use a different probability (here denoted as $\mathscr{Pr}$) which can be expressed in terms of the Bayesian Pr, in the following way: $\mathscr{Pr}(\cdot) := \mathscr{Pr}^{\theta}_{h}(\cdot) = \Pr(\cdot \mid \Theta = \theta, H = h)$. Thus, the frequentist likelihood ratio (denoted as $\mathscr{LR}$) is defined as

$$\mathscr{LR} = \frac{\mathscr{Pr}^{\theta}_{h_p}(E = e, B = b)}{\mathscr{Pr}^{\theta}_{h_d}(E = e, B = b)} = \frac{\mathscr{Pr}^{\theta}_{h_p}(E = e \mid B = b)}{\mathscr{Pr}^{\theta}_{h_d}(E = e \mid B = b)} = \frac{\Pr(E = e \mid B = b, \Theta = \theta, H = h_p)}{\Pr(E = e \mid B = b, \Theta = \theta, H = h_d)}.$$

Depending on the preferences of the expert, frequentist or Bayesian likelihood ratios can be used for the evaluation of forensic data. Once a choice has been made, it is important to be consistent with it, but literature often mixes up the two. At the best of our knowledge, this paper and (Cereda, 2016b) constitute the only forensic literature discussing the differences between the two approaches. (Cereda, 2016b) is concerned with the theoretical foundations of frequentist solutions, while this paper provides a careful derivation of the proper Bayesian LR for the rare type match problem described in Section 6.3: the situation in which the Y-STR profile of the crime stain and that of the suspect match but they are not among the Y-STR profiles observed in the reference database. In Section 6.4 we will discuss the fact that

influential Bayesian forensic literature (Weir, 1996; Aitken and Taroni, 2004; Taroni et al., 2010, 2014; Sjerps et al., 2016) seems to suggest the use of a frequentist defined likelihood ratio ($\mathcal{LR}$). They use Bayesian methodologies only inasmuch they provide a Bayesian estimate of θ to be plugged into $\mathcal{LR}$. Others (Curran, 2005; Van der Hout and Alberink, 2015), treat the likelihood ratio as function of θ and provide its posterior distribution with respect to the posterior distribution of θ given the data. However, one of the main points of discussion is that there is no need of such hybrid derivations, since the proper Bayesian LR is often very easy to obtain: this paper shows how this should be done, taking advantage of a very useful Lemma, presented in Section 6.5. However, for this method to be advisable, the Bayesian prior should be chosen in a sensible way, reflecting the expert's opinion, and not by mathematical convenience as it often happens.

The two most common Bayesian models (beta-binomial and Dirichlet-multinomial) are discussed in Sections 6.6 and 6.7. They are general enough to be applied to different kinds of forensic evidence evaluation, but in this paper they are applied to Y-STR profiles, with the double aim of exploring the performance of the conventional Bayesian prior choices for the rare type match case for non autosomal DNA profiles, and of showing how a full Bayesian LR is to be defined and calculated. Sensitivity analysis and a comparison with proposed hybrid plug-in solutions are carried out. We are not entirely satisfied with the performance of classical models for the rare haplotype match problem, which we believe would need different kinds of prior, more realistic and tuneable, such as those proposed in Cereda (2016c).

6.1.1 Notation

Throughout the paper the following notation is chosen: random variables and their values are denoted, respectively, with uppercase and lowercase characters: x is a specific realisation of X. Random vectors and their values are denoted, respectively, by uppercase and lowercase bold characters: $\mathbf{p}$ is a realisation of the random vector $\mathbf{P}$. Bayesian probability is denoted with $\Pr(\cdot)$, while the density of a continuous random variable X is denoted by $p(x)$. For a discrete random variable Y, both the continuous notation $p(y)$ and the discrete one $\Pr(Y = y)$ will be used, when there is no possibility of confusion. Frequentist probability will be denoted as $\mathcal{Pr}$.

6.2 Genetic terminology

The DNA sequence of an individual is a very long sequence of four letters (A, C, T, and G, each corresponding to four nucleotides) which code the genetic instructions necessary for the life of the individual. The entire sequence is unique to each individual (with the only exception of monozygotic twins, which share the same sequence), but DNA profiles used for forensic identification only describe a limited number of portions of this long sequence, called *markers* or *loci*.

STR markers, short for 'short tandem repeat', are loci where patterns of two or more letters (such as AGGT) are repeated adjacent one another. The number of times the pattern is

repeated at a specific locus varies among individuals, and constitutes the so-called *STR allele* at that locus. Y-STR profiles (also called *haplotype*) are made of a short list of STR alleles located at a particular collection of loci (typically 7 to 23) on the Y chromosome (Gill et al., 2001). The Y-STR profile is passed down identical from father to son, and the variability present in nature is only due to mutations. As a result, a typical Y-STR database contains some haplotypes observed many times, and many observed few times. Sometimes, the quality of a recovered DNA stain is poor. This implies that we cannot infer the alleles at all loci. This kind of profiles are called "incomplete", but we ignore this possibility for the rest of the paper.

6.3 The rare type match problem

In order to evaluate a match between a recovered stain and a suspect Y-STR profile, we need to weigh how probable the observed match is under the hypothesis that the suspect left the stain against how probable is the match under the hypothesis that someone else left the stain. Clearly, assuming that a match is always detected correctly, the first probability is 1, and the second depends on the proportion θ of the profile in the relevant population.

To assess this proportion, the expert is usually given a list of profiles from a sample of individuals belonging to the relevant population. Problems arise when the observed frequency of this characteristic is 0, the so-called 'rare type match problem'. This problem, so substantial that it has been defined "the fundamental problem of forensic mathematics" by Brenner (2010), is particularly significant for 'new' types of forensic evidence is involved, where the size of the available database is still limited. This is the case, for instance, for DIP-STR markers (Cereda et al., 2014a)). The same happens when Y-chromosome or mitochondrial (Carracedo et al., 2000) DNA profiles are used, since the set of possible haplotypes is extremely large, and the coverage of available databases is often limited. The case of Y-STR DNA will thus be retained here as an extreme but in practice common and important way in which the problem of assessing the evidential value of a rare type match can arise. This is a very appropriate and paradigmatic example, since literature provides examples of different approaches to evaluate the evidential value of a rare Y-STR profile match (Roewer et al., 2000; Andersen et al., 2013b, e.g.), even though, in our opinion, a proper Bayesian derivation for the LR in the rare type match case hasn't been proposed yet.

We will now review some of the methods proposed by literature to address the problem of assessing the frequency of a type with zero occurrence, sometimes under the name of 'zero numerator problem' (e.g. Winkler et al., 2002). Notice that this is related, but not equivalent, to the problem of assessing the likelihood ratio in case of a rare type match.

The *empirical frequency estimator*, also called *naive estimator*, that uses the frequency of the characteristic in the database, puts unit probability mass on the set of already observed characteristics, and it is thus unprepared for the observation of a new type. A solution could be the *add-constant* estimators (in particular the well-known *add-one* estimator, due to Laplace (1814), and the *add-half* estimator of Krichevsky and Trofimov (1981)), which add a constant to the count of each type, included the unseen ones. However, these methods require to know the number of possible unseen types, and does not perform well when this number is large

compared to the sample size (see Gale and Church (1994) for additional discussion). Louis (1981) proposes the so-called 'rule of three', that states that if n is the size of the database, $3/n$ is a good approximation of the 95% upper bound for the frequency. This is also proposed in a Bayesian framework, by Jovanovic and Levy (1997); Winkler et al. (2002); Chen and McGee (2008). Alternatively, Good (1953), based on an intuition of A.M. Turing, proposed the nonparametric *Good Turing estimator* for the total unobserved probability mass, based on the proportion of singleton observations in the sample. An extension of this estimator is applied to the LR assessment in the rare type match in Cereda (2016b). For a comparison between *add one* and *Good-Turing* estimator, see Orlitsky et al. (2003). As pointed out in Anevski et al. (2013), the *naive estimator*, and the *Good Turing estimator* are in some sense complementary: the first gives a good estimate for the observed types, and the second for the probability mass of the unobserved ones. More recently, Orlitsky et al. (2004) introduced the *high profile estimator*, which extends the tail of the *naive estimator* to the region of unobserved types. Anevski et al. (2013) improved this estimator and provided the consistency proof. Papers that address the rare Y-STR haplotype problem in forensic context are for instance Egeland and Salas (2008), Brenner (2010), Cereda (2016b), and Cereda (2016c). Moreover, the Discrete Laplace method presented in Andersen et al. (2013b), even though not specifically designed for the rare type match, can be successfully applied to that case (Cereda, 2016b).

Bayesian nonparametric estimators for the probability of observing a new type have been proposed by, e.g., Tiwari and Tripathi (1989), Lijoi et al. (2007), and Favaro et al. (2009). However, for the likelihood ratio assessment not only the probability of observing a new species is required but also the probability of observing this same species twice (according to the defence, the profile of the crime stain and of the suspect are two independent observations). Cereda (2016c) is the first paper that addresses the problem of LR assessment in the rare haplotype match case using Bayesian nonparametric models.

6.4 The full Bayesian approach to LR

The likelihood ratio assessment often involves some unknown nuisance parameters, denoted as θ. In our case, it is the proportion of individuals of the relevant population with Y-STR profile corresponding to that of the matching trace, or the entire vector containing the population proportions of all the DNA profiles. The parameter of interest, h, is the unknown true hypothesis. Available data is made of evidence (E) directly related to the crime, which helps to discriminate h, and additional background data (B) not directly related to the crime and only pertaining to the nuisance parameter θ. This is partially different from the 'background information' I as defined in Aitken and Taroni (2004), and Taroni et al. (2014), but often background data can be thought of as part of the background information.

The difference between Bayesian and frequentist methods consists in how they treat the parameters θ and h. A Bayesian statistician models the uncertainty about their value by random variables Θ and H, which are given prior distributions $p(\theta)$ and $p(h)$. Frequentists consider them as fixed (i.e., without distribution) unknown quantities. The reader is invited to notice the difference between θ and h: one is the parameter which we 'test' through the

likelihood ratio (h), the other (θ) is a nuisance parameter involved in the calculation of the LR. Some assumptions about the conditional independence probability for the model can be made, valid both for the frequentist and for the Bayesian approach:

a. The distribution of B given h and θ, only depends on θ.

b. B is independent of E, given θ and h.

In our DNA example, condition **a** corresponds to ask that the sampling mechanism to obtain the database of reference is independent of which hypothesis is correct. This is true if the database is collected before the crime.

Condition **b** holds if the suspect has been found based on different evidence, i.e. not the result of for example, a DNA database search. In what follows, we are going to use Bayesian networks notation to specify the conditional independence relations of the proposed models. We expect the reader to be familiar with such a representation, if not we suggest to read Koski and Noble (2011).

6.4.1 Bayesian point of view.

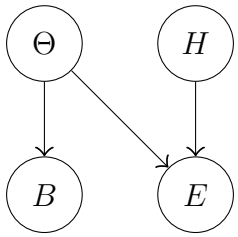


Figure 6.1: Bayesian network representing the dependency relations between E (evidence of the case) B (background data in the form of a database) Θ (population parameter) and H (hypotheses of interest).

Bayesians deal with the uncertainty over the parameters θ and h by considering their values as realisations of, respectively, random variables Θ and H. A full Bayesian model is defined when the prior joint probability distribution for all the random variables of the model (here E, B, H and Θ) is given. This full Bayesian model can be thus represented by the Bayesian network of Figure 6.1, which is in turn equivalent to the following three conditions:

Bayesian a. B is conditionally independent of H given Θ.

Bayesian b. B is conditionally independent of E given Θ and H.

Bayesian c. Θ is unconditionally independent of H.

Notice that they are the Bayesian reformulation of conditions **a.** and **b.** mentioned above, with an additional condition (**Bayesian c.**) which corresponds to assuming that the Bayesian probability makes Θ and H independent, and that is guaranteed for instance if prior beliefs on θ and on h are assessed by people with different responsibilities and tasks: a judge for h and a DNA expert (or a statistician) for θ. However, notice that by definition the LR is independent of the prior belief over h.

The structure of the Bayesian network (or, equivalently, the three conditions above) allows to factorise the joint prior as $p(\theta, h, b, e) = p(\theta)p(h)p(b|\theta)p(e|\theta, h)$. The Bayesian probability Pr underlying to the model is defined accordingly. Like all Bayesian probabilities it is an expression of the subjective belief of the experts. This is achieved by choosing the prior distribution for θ and h reflecting the expert's beliefs. The distribution of all other variables given θ and h is defined by the model, and needs no subjective assessment.

The Bayesian likelihood ratio can be derived in the following way:

$$\begin{aligned}\mathrm{LR} =& \frac{\Pr(E = e, B = b|H = h_p)}{\Pr(E = e, B = b|H = h_d)} = \frac{\Pr(E = e|B = b, H = h_p)}{\Pr(E = e|B = b, H = h_d)} = \frac{\int p(e|b, h_p, \theta)\, p(\theta|b, h_p)\mathrm{d}\theta}{\int p(e|b, h_d, \theta)p(\theta|b, h_d)\mathrm{d}\theta} \\ =& \frac{\int p(e|h_p, \theta)\, p(\theta|b)\, \mathrm{d}\theta}{\int p(e|h_d, \theta)\, p(\theta|b)\, \mathrm{d}\theta} = \frac{\mathbb{E}(\Pr(E = e|H = h_p, \Theta) \mid B = b)}{\mathbb{E}(\Pr(E = e|H = h_d, \Theta) \mid B = b)},\end{aligned} \tag{6.2}$$

where the second equality is due to the independence of B and H, and the fourth one both to condition **b** and to the independence of Θ and H given B. These independence properties follow from the network's structure.

6.4.2 Frequentist point of view.

As already mentioned, frequentists consider h and θ as fixed quantities, whose unknown values correspond to, respectively, the true value of θ and the correct hypothesis. The frequentist model can be thus seen as a special case of the Bayesian model described in Section 6.4.1, where Θ and H are given degenerate priors on θ and h, respectively. Alternatively, one can express the frequentist probability $\mathcal{Pr}$ in terms of the Bayesian Pr in the following way: $\mathcal{Pr}(\cdot) := \mathcal{Pr}^{\theta}_{h}(\cdot) = \Pr(\cdot|H = h, \Theta = \theta)$. If the Bayesian Pr was subjective, the frequentist $\mathcal{Pr}$ is a measure which is universally determined by nature. Regarding h, according to prosecution its true value is h_p, while according to defence it is h_d. So one can think of two different frequentist probabilities: one for the prosecution ($\mathcal{Pr}^{\theta}_{h_p}$) and one for the defence ($\mathcal{Pr}^{\theta}_{h_d}$). From a frequentist point of view, conditions **a** and **b** correspond to ask that:

Frequentist a. $\mathcal{Pr}^{\theta}_{h_p}(B = b) = \mathcal{Pr}^{\theta}_{h_d}(B = b)$, for all θ and b.

Frequentist b. $\mathcal{Pr}^{\theta}_{h}(E = e|B = b) = \mathcal{Pr}^{\theta}_{h}(E = e)$, for all θ, h, e, and b.

Obviously, **Bayesian c** becomes irrelevant in the frequentist framework. The frequentist $\mathcal{LR}$ can be derived as:

$$\mathcal{LR} = \frac{\mathcal{Pr}^{\theta}_{h_p}(E = e, B = b)}{\mathcal{Pr}^{\theta}_{h_d}(E = e, B = b)} = \frac{\mathcal{Pr}^{\theta}_{h_p}(E = e \mid B = b)\mathcal{Pr}^{\theta}_{h_p}(B = b)}{\mathcal{Pr}^{\theta}_{h_d}(E = e \mid B = b)\mathcal{Pr}^{\theta}_{h_d}(B = b)} = \frac{\mathcal{Pr}^{\theta}_{h_p}(E = e)}{\mathcal{Pr}^{\theta}_{h_d}(E = e)} \tag{6.3}$$

where the last equality is due to conditions **Frequentist a**, and **Frequentist b**.

Stated otherwise, frequentists look at a value for $\mathrm{LR}|\theta$ (read "LR given θ"), where the value θ is fixed and has to be estimated through data. Through observations, frequentists attempt to get close to the true $\mathcal{LR}$ by choosing some estimator $\widehat{\mathcal{LR}}$. One possibility is to estimate θ with a particular $\widehat{\theta}$. This leads to the so-called *plug-in estimation* $\widehat{\mathcal{LR}} = \mathcal{LR}(\widehat{\theta})$ of the $\mathcal{LR}$.

However, that's not the only option (Cereda, 2016b). By looking at (6.3) the reader will realise that, if the frequentist approach is chosen, and under conditions **a** and **b**, one would get to the same result by evaluating only E or both E and B. This means that part of the information, namely B, is not useful to discriminate between the two hypotheses of interest (however, it usually plays an important role to obtain the estimate $\widehat{\theta}$ to be plugged into the $\mathcal{LR}$). The same does not hold in the Bayesian context.

6.4.3 The Bayesian plug-in LR and the proper Bayesian LR

It is now time to discuss the fact that important forensic literature (e.g. Evett and Weir, 1998; Balding, 2005; Lucy, 2005) considers the likelihood ratio as 'a measure of the probative value of the evidence regarding the two hypotheses' h_p and h_d. According to this, it indicates the extent to which E (and only E) is in favour of one hypothesis over the other. This is, in our opinion, the first important problem, since all data at disposal (namely E and B) should be evaluated. Even though this is irrelevant in the frequentist framework (see (6.3)), in the Bayesian framework for this definition to be appropriate one needs to replace the probability Pr with the posterior probability $\Pr^*(\cdot) = \Pr(\cdot \mid B = b)$. Indeed, it holds that

$$\mathrm{LR} = \frac{\Pr(E = e, B = b \mid H = h_p)}{\Pr(E = e, B = b \mid H = h_d)} = \frac{\Pr(E = e \mid B = b, H = h_p)}{\Pr(E = e \mid B = b, H = h_d)} = \frac{\Pr^*(E = e \mid H = h_p)}{\Pr^*(E = e \mid H = h_d)}.$$

It is as if we have splitted the evaluation process in two steps: first we observe $B = b$, and update the probability $\Pr(\cdot)$ to the posterior $\Pr^*(\cdot) = \Pr(\cdot \mid B = b)$, and then we define the likelihood ratio as the ratio of the probabilities ($\Pr^*$) of observing (only) the evidence E, under the two alternative hypotheses.[1] With the exception of little literature (e.g., Dawid and Mortera (1996); Brümmer and Swart (2014); Taroni et al. (2016)), this point is generally mistaken and the problem is split into two phases. A Bayesian estimate of θ using B, in the form of a posterior expectation, is obtained, and then plugged into a frequentist defined $\mathcal{LR}$. It is as if, instead of using a combined model such as that in Figure 6.1, two separate models as those in Figure 6.2 are used: the left one is used to update the prior over the parameter, while the second one is used to derive the likelihood ratio (with θ considered as a fixed quantity). In the end, θ is replaced with the posterior expectation of Θ given B. This method will be referred to in the paper as the 'Bayesian plug-in method', since it is wrongly considered Bayesian, but it actually plugs-in Bayes estimates into a likelihood ratio defined in a frequentist way. The correct Bayesian approach would be either to evaluate both E and B simultaneously, using the network of Figure 6.1, or in two steps: after the observation of B, we can update the model to the one represented in Figure 6.3, and use this for the evaluation of E.

6.4.4 State of the art for DNA match evaluation

In case of a DNA match, we can use the Bayesian network of Figure 6.4, which is equivalent to the network in Figure 6.1 with the only difference that here node E is split into two separate

[1] Often, in literature (Taroni et al., 2014, e.g.), it is explicitly stated that I, the so-called background information, is omitted in the notation. We then agree with this choice provided that B is part of I.

Figure 6.2: The two phase approach corresponding to Bayesian plug-in.

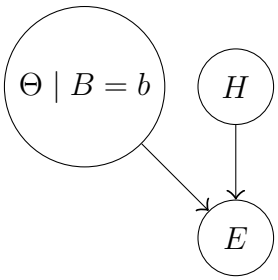


Figure 6.3: Updated Bayesian network after the observation of B.

nodes, E_s and E_c representing the suspect's and the crime stain's profile, respectively. We denote with $\boldsymbol{\theta}$ the unknown vector made of the population proportions of the different Y-STR profiles in the relevant population, modelled through the random variable $\boldsymbol{\Theta}$. Here, we assume that we know the whole list of different DNA types present in the relevant population, while later we will consider the situation in which we don't. With Θ_{e_s} we will denote the population frequency of the suspect's (and crime stain's) profile.

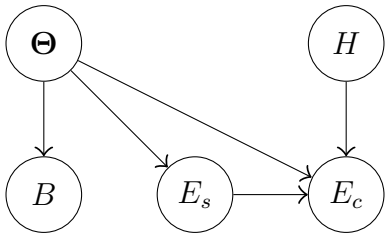


Figure 6.4: Bayesian network for the DNA example.

According to the prosecution, the suspect left the stain. This implies that $\Pr(E_c = e_s|\boldsymbol{\Theta} = \boldsymbol{\theta}, E_s = e_s, H = h_p) = 1$, under the assumption that each true match is correctly reported. According to the defence, another person from the population left the stain, hence the probability of it being exactly of type e_c is equal to the population proportion of that profile: $\Pr(E_c = e_s|\boldsymbol{\Theta} = \boldsymbol{\theta}, E_s = e_s, H = h_d) = \theta_{e_s}$. Moreover, it holds that $p(b|e_s, \boldsymbol{\theta})p(e_s|\boldsymbol{\theta})p(\boldsymbol{\theta})$ is proportional to $p(\boldsymbol{\theta}|e_s, b)$. The correct Bayesian procedure would lead to:

$$\begin{aligned}\mathrm{LR} =& \frac{\Pr(E = e, B = b|H = h_p)}{\Pr(E = e, B = b|H = h_d)} = \frac{\int \Pr(E_c = e_s|H = h_p, \boldsymbol{\Theta} = \boldsymbol{\theta}, E_s = e_s)p(e_s|\boldsymbol{\theta})p(b|\boldsymbol{\theta}, e_s)p(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int \Pr(E_c = e_s|H = h_d, \boldsymbol{\Theta} = \boldsymbol{\theta}, E_s = e_s)p(e_s|\boldsymbol{\theta})p(b|\boldsymbol{\theta}, e_s)p(\boldsymbol{\theta})d\boldsymbol{\theta}} \\ =& \frac{\int p(\boldsymbol{\theta}|e_s, b)d\theta}{\int \theta_{e_s} p(\boldsymbol{\theta}|e_s, b)d\theta} = \frac{1}{\mathbb{E}(\Theta_{e_s}|E_s = e_s, B = b)}.\end{aligned}$$

On the other hand, the common approach taken by forensic literature would be to compute the likelihood ratio as if θ was given. Given θ, B is conditionally independent of the rest of the variables, hence it can be removed by the formula.

$$\text{LR} = \frac{\Pr(E = e|\Theta = \theta_{e_s}, H = h_p)}{\Pr(E = e|\Theta = \theta_{e_s}, H = h_d)} = \frac{\Pr(E_c = e_s|\Theta = \theta_{e_s}, E_s = e_s, H = h_p)}{\Pr(E_c = e_s|\Theta = \theta_{e_s}, E_s = e_s, H = h_d)} = \frac{1}{\theta_{e_s}}.$$

Then, θ_{e_s} is replaced with $\widehat{\theta_{e_s}} = \mathbb{E}(\Theta_{e_s}|B = b)$. In the end, computationally, the difference amounts on using $\mathbb{E}(\Theta_{e_s}|B = b, E_s = e_s)$ instead of $\mathbb{E}(\Theta_{e_s}|B = b)$ (i.e., the well-known problem of whether to add or not the suspect to the database before taking the posterior) and thus the plug-in can be seen as an approximation of the full Bayesian approach. However, it is an hybrid solution, thus conceptually ill-defined. This hybrid approach is often considered Bayesian since the lack of knowledge about θ is dealt with using the Bayesian posterior mean $\widehat{\theta_{e_s}} = \mathbb{E}(\Theta_{e_s}|B)$ as a point estimate of θ_{e_s} (Weir, 1996; Curran, 2005; Taroni et al., 2010; Sjerps et al., 2016). This is why we will refer to this way of proceeding as the 'Bayesian plug-in method'. As pointed out in Weir (1996), "either the mean or the mode of the posterior distribution can serve as an estimate but each is merely a summary of the whole distribution". Not only this method is hybrid and inconsistent, but it suffers from several weaknesses. For instance, one would obtain different $\widehat{\mathcal{LR}}$s depending on whether one wants to estimate θ_{e_s}, $1/\theta_{e_s}$ or $\log_{10}(1/\theta_{e_s})$: this arbitrariness is in some way entailed in the idea of 'estimating' the likelihood ratio. Moreover, as stated in Taroni et al. (2016), the likelihood ratio (meaning the Bayesian one) should be calculated, rather than estimated. Including B as part of the data to evaluate, and applying the Bayesian theory, we can calculate the Bayesian LR, without any estimation needed. Notice that already Foreman et al. (1997) and Brümmer and Swart (2014) proposed a differentiation between the 'plug-in estimates' and the 'full Bayesian analysis'.

6.5 A useful Lemma

Lemma 2 is a result regarding four general random variables A, X, Y, H whose conditional dependencies are represented by the Bayesian network of Figure 6.5. This lemma is important due to the possibility of applying it to a very common forensic situation: the prosecution and the defence disagree on the distribution of part of data (Y) but agree on the distribution of the other part (X). The distribution of X and Y depends on some parameter(s) modelled by A.

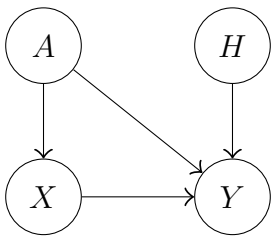


Figure 6.5: Conditional dependencies of the random variables of Lemma 2

Lemma 1. *Given four random variables A, H, X and Y, whose conditional dependencies are represented by the Bayesian network of Figure 6.5, the likelihood function for h, given $X = x$ and $Y = y$ satisfies*

$$\text{lik}(h \mid x, y) \propto \mathbb{E}(p(y \mid x, A, h) \mid X = x). \tag{6.4}$$

For given x, y and h, $p(y \mid x, A, h)$ in the right hand side of (6.4) is a function of the random quantity A. The expectation is taken with respect to the conditional distribution of A given $X = x$. A proof of this lemma can be found in Cereda (2016b). We will see an application of it in Sections 6.6 and 6.7.

6.6 Bayesian LR calculation, based on beta-binomial model

In the binomial model, the database of size N is regarded as the result of a sequence of N Bernoulli trials with parameter θ, where *success* corresponds to the observation of the same haplotype as that observed at the crime scene, and failure to the observation of any other type. Let's denote by b the number of successes among these N experiments. When data is treated as a binomial outcome, the most conventional choice for the prior over the parameter θ (probability of success) is the beta distribution, due to the famous conjugacy property. In forensic and medical statistics literature there are many examples of the use of this distribution for a genetic (autosomal) proportion (Weir, 1996; Gunel and Wearden, 1995; Roewer et al., 2000; Brenner, 2010; Buckleton et al., 2011; Biedermann et al., 2008, 2013).

$$\Theta \sim \text{Beta}(\alpha, \beta).$$

The observation of the suspect's profile E_s corresponds to another Bernoulli trial, a successful one in the case of interest (the suspect matches the crime stain type). The information provided by the database and the suspect's type can be reduced to the count of profiles of this type in this sample of size $N + 1$ (database and suspect) from the population of interest.

$$B \mid \Theta = \theta \sim \text{Bin}(N, \theta)$$

$$B, E_s \mid \Theta = \theta \sim \text{Bin}(N + 1, \theta)$$

Notice that according to the defence E_c can be seen as another Bernoulli experiment of the same kind. On the other hand, according to the prosecution it is equal to 1 with probability one. Stated otherwise,

$$E_c \mid E_s = 1, H = h \sim \begin{cases} \delta_1 & \text{if } H = h_p \\ \theta & \text{if } H = h_d \end{cases},$$

where δ represents the Dirac delta function. The Bayesian network of Figure 6.4 can be used for this model. Hence, we can apply the Lemma 2 using $X = (B, E_s)$ (the part of data whose distribution is agreed on by defence and prosecution) and $Y = E_c$ (the part of data whose

distribution is disagreed on by defence and prosecution). The LR can thus be developed in the following way:

$$\mathrm{LR} = \frac{\mathbb{E}(\Pr(E_c = 1 | E_s = 1, H = h_p, \Theta) \mid E_s = 1, B = b)}{\mathbb{E}(\Pr(E_c = 1 | E_s = 1, H = h_d, \Theta) \mid E_s = 1, B = b)} = \frac{1}{\mathbb{E}(\Theta \mid E_s = 1, B = b)} = \frac{\alpha + \beta + N + 1}{\alpha + b + 1}. \tag{6.5}$$

The last equality is due to the fact that, using the well-known beta-binomial conjugacy property, it holds that

$$\Theta \mid B = b, E_s = 1 \sim \mathrm{Beta}(\alpha + b + 1, \beta + N - b).$$

The LR as in (6.5), also proposed in Dawid and Mortera (1996) and Taroni et al. (2016), can be compared to the one obtained with the 'standard' Bayesian plug-in estimate (Weir, 1996; Taroni et al., 2010):

$$\widehat{\mathrm{LR}} = \frac{\alpha + \beta + N}{\alpha + b}.$$

It is easy to see that the Bayesian plug-in $\widehat{\mathrm{LR}}$ is a non-conservative estimate of LR, in a way that is unfavourable to the defence. Indeed, $\mathrm{LR} < \widehat{\mathrm{LR}} \Leftrightarrow \beta + N > b$, which is always true, since $b \leq N$ and $\beta > 0$. Notice that there is an alternative derivation for (6.5). It can be obtained in a two-step evaluation: first, the observation of the database B and of the suspect haplotype E_s updates the probability Pr, then the updated probability is used to calculate the likelihood ratio for the observation of another identical haplotype (the one found at the crime scene).

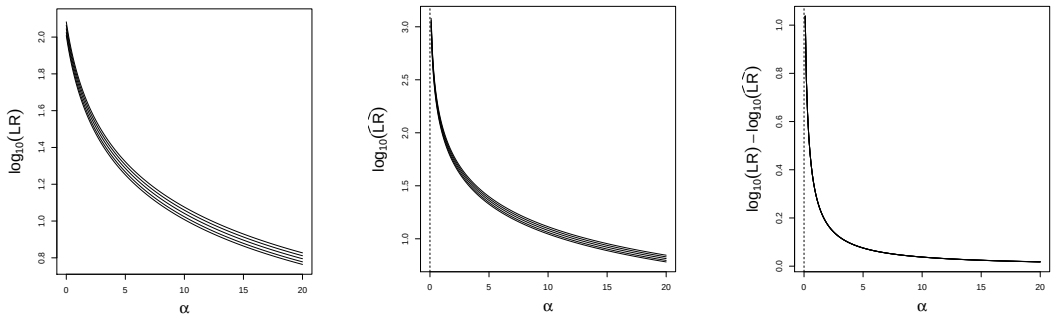
First step The probability Pr is updated to $\Pr^{**}(\cdot) = \Pr(\cdot | E_s = 1, B = b)$ after the database and the haplotype of the suspect are observed. In practice, the prior distribution $\mathrm{Beta}(\alpha, \beta)$ on θ is updated to the posterior $\mathrm{Beta}(\alpha + 1, \beta + N - 1)$.

Second step The new probability $\Pr^{**}$ is used to calculate the likelihood ratio for the observation of the haplotype from the crime scene:

$$\mathrm{LR} = \frac{\Pr^{**}(E_c = 1 | H = h_p)}{\Pr^{**}(E_c = 1 | H = h_d)} = \frac{1}{\mathbb{E}^{**}(\Theta)} = \frac{1}{\frac{\alpha+b+1}{\alpha+\beta+1+N}}.$$

Sensitivity analysis. The sensitivity of the quantities $\log_{10} \mathrm{LR}$, $\log_{10} \widehat{\mathrm{LR}}$, and of the difference between them, to the hyperparameters α and β of the beta prior is shown in Figure 6.6, for the rare type match case (i.e., $b = 0$), and with $N = 100$. In particular, the figure shows the variation of $\log_{10} \mathrm{LR}$ (a), of the plug-in estimate $\widehat{\log_{10} \mathrm{LR}} = \log_{10} \widehat{\mathrm{LR}}$ (b), and of the difference $\log_{10} \widehat{\mathrm{LR}} - \log_{10} \mathrm{LR}$ (c), when different values of α (x axis) and β (only five values corresponding to the different lines) are chosen in the interval $(0, 20]$.

Observing Figure 6.6 (or analysing (6.5)), it can be seen that the three quantities of interest hardly depend on β, while they decrease as α increases. In particular, when α decreases to 0, $\log_{10} \mathrm{LR}$ behaves as $\log_{10}(1 + \beta + N)$, while $\log_{10} \widehat{\mathrm{LR}}$ increases to $+\infty$. Another way to see this is that, for fixed β, as α increases, the prior distribution of θ resembles more and more



(a): $\log_{10} \mathrm{LR}$ (b): $\log_{10} \widehat{\mathrm{LR}}$ (c): $\log_{10} \widehat{\mathrm{LR}} - \log_{10} \mathrm{LR}$

Figure 6.6: Sensitivity analysis for the three quantities $\log_{10} \mathrm{LR}$ (a), $\log_{10} \widehat{\mathrm{LR}}$ (b), $\log_{10} \widehat{\mathrm{LR}} - \log_{10} \mathrm{LR}$ (c), in the beta-binomial model, when $\alpha \in (0, 20]$ (x axis), and for $\beta \in \{1, 5, 10, 15, 20\}$ (corresponding to the different lines, where highest line corresponds to highest β).

to the degenerate distribution localised on the value $\theta = 1$ (notice that this is inappropriate for the rare type match case). This means that the haplotype whose population proportion is modelled through the random variable Θ (i.e., the haplotype of the crime stain and of the suspect) has probability one to be observed, which leads to $\widehat{\mathrm{LR}} = 1$ (hence, $\log_{10} \mathrm{LR} = 0$). On the other hand, if α decreases to zero, the prior distribution over θ tends to resemble to the degenerate distribution localised on the value $\theta = 0$. This leads to $\widehat{\mathrm{LR}} = 1/0 = +\infty$. On the whole, the plug-in estimate $\widehat{\mathrm{LR}}$ is less stable than LR, as can be seen comparing Figures 6.6 (a) and (b), in the sense that is more sensitive to changes in α (especially for small values).

The difference, represented in (c) has, for fixed β, a vertical asymptote when $\alpha \to 0$, increasing as fast as $\log_{10} 1/\alpha$. On the other hand, it decreases to 0 with an horizontal asymptote when $\alpha \to \infty$. From Figure 6.6 (c) it can be observed that the difference is important only for small values of α. Otherwise the two methods would lead essentially to the same conclusions, so that the plug-in can be seen as a good approximation of the proper Bayesian procedure.

6.7 Bayesian LR calculation, based on Dirichlet-multinomial model

When the database is treated as a multinomial sample of size N from a population with k different haplotypes, the conventional choice for the prior over the vector $\boldsymbol{\theta}$ containing the population proportions of all the different haplotypes in nature is the Dirichlet distribution. Literature provides many examples of the use of this prior for the population proportions of autosomal markers (Curran et al., 2002; Balding, 1995; Lange, 1995; Weir, 1996; Buckleton and Curran, 2005; Taroni et al., 2010, e.g.). However, these examples don't consider the uncertainty about the number k of possible types in the population, and this can be a

problem especially since we want to apply it to Y-STR haplotypes, for which the database often does not offer a good coverage. If, in addition, we are using the model for the rare type match case, then we have to find a solution.

The problem of estimating k is a very challenging one. It has been addressed both with frequentist methods (Chao and Lee, 1992; Haas and Stokes, 1998, e.g.) and with Bayesian methods (Hill, 1968; Lewins and Joanes, 1984; Barger and Bunge, 2010, e.g.). We propose the derivation of a full Bayesian LR which models the uncertainty over the number k of different types in the population, with priors . The model is represented by the Bayesian network of Figure 6.7. The bottom part (from node $\mathbf{\Theta}$ down) has a well-known structure (see Figure 6.4), while the upper part needs further explanation.

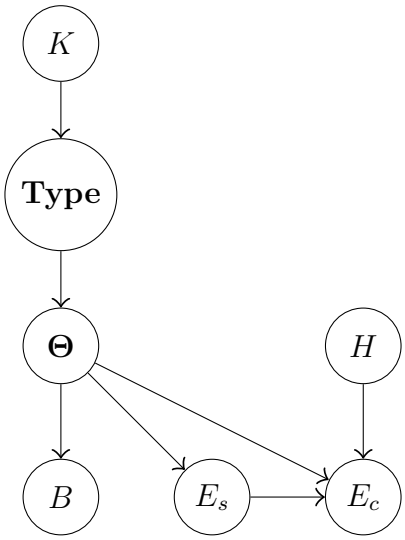


Figure 6.7: Bayesian network for Dirichlet-multinomial model, when k is randomized.

Assume that there may be at most m theoretically possible profiles, alphabetically[2] ordered in a vector, called $\mathbf{s}$. For instance, $m = 20^{10}$ (10 loci, with 20 possible alleles each). Only k of them are actually present in nature (or more specifically in the population of interest), but k is not known and also which of the m are those k is not known. We will denote as K the random variable which represents how many of the m possible haplotypes are actually present in the population of interest. The prior distribution for k is denoted generically as $p(k)$. The random vector **Type**, of length k, contains the ordered positions, in the vector $\mathbf{s}$, of the k haplotypes of the population of interest. A particular configuration of **Type** is denoted as $\mathbf{t} = (i_1, ..., i_k)$, where $i_1 < ... < i_k$. $\mathbf{t}$ is chosen uniformly at random from the possible $\binom{m}{k}$ configurations. The random vector $\mathbf{\Theta}$ contains the population proportions of all the haplotypes, both those whose position is contained in **Type**, and those that are not (corresponding to zero entries). A particular configuration of $\mathbf{\Theta}$ is denoted as $\boldsymbol{\theta} = (\theta_1, ..., \theta_m)$, many entries of which are zero. We assume that the positive entries, i.e., $(\theta_i \mid i \in \mathbf{t})$, are drawn from a k dimensional Dirichlet distribution with all the k hyperparameters α equal to 1. Now, as usual, H represents the hypotheses of interest, and can take the value $h \in \{h_p, h_d\}$,

[2]Remember each profile is a list of numbers.

according to the prosecution or the defence, respectively. E_s and E_c contain the index e_s and e_c of the haplotypes of the suspect and of the crime scene, respectively. In the situation of interest $e_c = e_s$. Lastly, the random vector $\mathbf{B}$ represents the database, seen as a multinomial sample from the population with parameters N and $\boldsymbol{\theta}$. A particular configuration of $\mathbf{B}$ is denoted as $\mathbf{b} = (b_1, ..., b_m)$ representing the absolute frequency in the database of each of the m haplotypes. It contains $k_{obs} < k$ positive values, and many zeros. By applying Lemma 2 to this situation we have that

$$\mathrm{LR} = \frac{\mathbb{E}(\Pr(E_c = e_s \mid E_s = e_s, \mathbf{B} = \mathbf{b}, \boldsymbol{\Theta}, H = h_p) \mid E_s = e_s, \mathbf{B} = \mathbf{b})}{\mathbb{E}(\Pr(E_c = e_s \mid E_s = e_s, \mathbf{B} = \mathbf{b}, \boldsymbol{\Theta}, H = h_d) \mid E_s = e_s, \mathbf{B} = \mathbf{b})} = \frac{1}{\mathbb{E}(\Theta_{e_s} \mid E_s = e_s, \mathbf{B} = \mathbf{b})}. \tag{6.6}$$

It can be shown that for $\alpha = 1$ this leads to

$$\mathrm{LR} = \frac{1}{2} \frac{\displaystyle\sum_{k=k_{obs}+1}^{m} \binom{k}{k_{obs}+1} \frac{\Gamma(k)}{\Gamma(k+N+1)} p(k)}{\displaystyle\sum_{k=k_{obs}+1}^{m} \binom{k}{k_{obs}+1} \frac{\Gamma(k)}{\Gamma(k+N+2)} p(k)}. \tag{6.7}$$

Interested readers can refer to the Appendix of Cereda et al. (2016) for the complete derivation of (6.7), outlined as follows: first, the denominator of the right hand side of (6.6) can be expressed as $\mathbb{E}(\mathbb{E}(\Theta_{e_s} \mid E_s = e_s, \mathbf{B} = \mathbf{b}, K) \mid \mathbf{B} = \mathbf{b}, E_s = e_s)$, where the outside expectation is taken with respect to K given $\mathbf{B} = \mathbf{b}, E_s = e_s$. Next, in the mentioned appendix, the conditional distributions needed to compute this double expectation, namely that of $\boldsymbol{\Theta}$ given $\mathbf{B} = \mathbf{b}, E_s = e_s, K = k$, and that of K given $\mathbf{B} = \mathbf{b}, E_s = e_s$, are derived.

Notice that the likelihood ratio depends on the data only through k_{obs}. This is due to the choice of the symmetric Dirichlet prior, and of the uniform prior for **Type**. In particular, this tells us that data can be reduced by sufficiency to k_{obs}. The likelihood ratio obtained through a classical plug-in Bayesian estimation is:

$$\widehat{\mathrm{LR}} = \frac{\bar{k}\alpha + N}{\alpha + b_{e_s}} = \bar{k} + N, \tag{6.8}$$

where the number of haplotypes is a fixed value $\bar{k}$, to be chosen (or estimated) in advance. In order to compare the two values (6.7) and (6.8), we need to choose a value for $\bar{k}$. A reasonable choice can be $\bar{k} = \mathbb{E}(K)$. Among the possible priors over K, we decided to test the Poisson distribution (see Section 6.7.1) and the negative binomial distribution (see Section 6.7.2).

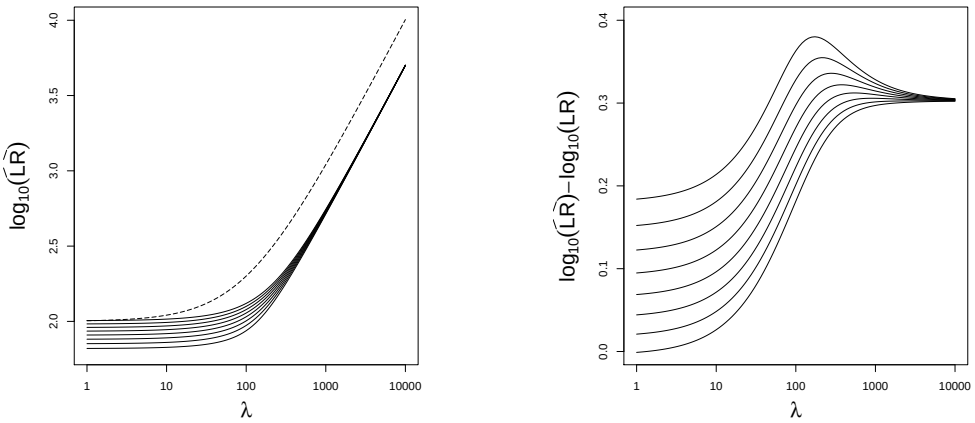
6.7.1 Poisson prior

In this section a Poisson distribution with parameter λ, truncated so as to have support only on $\{1, 2, ..., m\}$, is chosen as prior distribution for K. If λ and m are large enough, the normalising constant can be omitted and we have the standard Poisson distribution: The LR in (6.7) becomes

$$\mathrm{LR} = \frac{1}{2}\frac{\sum_{k=k_{obs}+1}^{m} \frac{\lambda^k}{k-k_{obs}-1!}\frac{\Gamma(k)}{\Gamma(N+k+1)}}{\sum_{k=k_{obs}+1}^{m} \frac{\lambda^k}{k-k_{obs}-1!}\frac{\Gamma(k)}{\Gamma(N+k+2)}}$$

It is then of interest to analyse the quantities $\log_{10}\mathrm{LR}$, $\log_{10}\widehat{\mathrm{LR}}$, and the difference $\log_{10}\widehat{\mathrm{LR}} - \log_{10}\mathrm{LR}$ between them and to carry on a sensitivity analysis to see how these quantities vary when the parameter λ changes.

Sensitivity analysis In the rare type match problem (i.e., $b_{e_s} = 0$), when a Poisson(λ) prior is chosen for the dimension K of the Dirichlet distribution (with all parameters α equal to 1), the sensitivity of the three quantities $\log_{10}\mathrm{LR}$, $\log_{10}\widehat{\mathrm{LR}}$, and of their difference, to λ and k_{obs}, is shown in Figure 6.8 for $N = 100$.



(a): $\log_{10}\widehat{\mathrm{LR}}$ (b): $\log_{10}\widehat{\mathrm{LR}} - \log_{10}\mathrm{LR}$

Figure 6.8: Poisson prior, for the Dirichlet model. Sensitivity analysis of $\log_{10}\mathrm{LR}$ ((a), black lines), $\log_{10}\widehat{\mathrm{LR}}$ ((a), dashed line), and of the difference $\log_{10}\widehat{\mathrm{LR}} - \log_{10}\mathrm{LR}$ (b), to different values of $\lambda \in [1, 10\,000]$ (x axis), and $k_{obs} \in \{30, 40, 50, 60, 70, 80, 90, 100\}$ (represented by the different lines where highest line corresponds to highest k_{obs}).

In particular, it can be inferred that the LR depends little on k_{obs} and a lot on λ. When λ is big (which is typically true λ being the expected value of the number of different Y-STR haplotypes in a population) the LR depends almost only on λ. In particular, LR increases linearly with λ, since $\mathrm{LR} \sim \lambda/2$. This can be explained by replacing the Poisson prior on k, by the degenerate distribution localised on (the integer part of) λ: $f_K(k) = f(k;\lambda) = \mathbb{1}_{\{\lambda\}}(k)$, for $\lambda \in \{1, 2, ...\}$. This approximation is sensible for large values of λ in virtue of the law of the large numbers (the Poisson(λ) being the sum of λ Poisson(1)). In this case (6.7) becomes

$$\mathrm{LR} = \frac{1+N+\lambda}{2} \sim \frac{\lambda}{2}, \text{ for } \lambda \to +\infty, \text{and } N \text{ fix.}$$

The plug-in estimates of $\log_{10}\widehat{\mathrm{LR}}$ (as defined in (6.8) and with the choice of $\bar{k} = \lambda$) is the dashed line shown in Figure 6.8 (a). The difference between the 'true' value $\log_{10}\mathrm{LR}$, and the estimated one $\log_{10}\widehat{\mathrm{LR}}$ is shown in Figures 6.8 (b). In particular, one can see that, for big λ it decreases when λ increases and depends a little on k_{obs}, while for small values of λ it has the opposite behaviour, and depends more strongly on k_{obs}. Note that, again, the plug-in method overestimates the LR by up to almost half of an order of magnitude.

6.7.2 Negative binomial prior

A different choice is that of using as prior for k the negative binomial distribution (Hill, 1968, 1979; Lewins and Joanes, 1984). For our model a negative binomial distribution truncated so as to have support $\{1, ..., m\}$ is more appropriate. If $\mathbb{E}(K)$ is large, but small compared to m, the standard negative binomial distribution can be used:

$$\Pr(K = k|r, q) = \binom{k+r-1}{k}(1-q)^k q^r, \quad \forall k \in \mathbb{N}.$$

where $r > 0$ and $q \in (0, 1)$. Using this prior, the likelihood ratio in (6.7) becomes:

$$\mathrm{LR} = \frac{1}{2}\frac{\sum_{k=k_{obs}+1}^{m}(1-q)^k \frac{\Gamma(k)}{\Gamma(k+N+1)}\frac{\Gamma(k+r)}{\Gamma(k-k_{obs})}}{\sum_{k=k_{obs}+1}^{m}(1-q)^k \frac{\Gamma(k)}{\Gamma(k+N+2)}\frac{\Gamma(k+r)}{\Gamma(k-k_{obs})}}. \tag{6.9}$$

In the following, a series of properties of the (zero truncated) negative binomial distribution will be listed, which help to understand why this choice is more appropriate than the choice of the Poisson distribution as a prior for K. We will denote as $\mathrm{NB}(r, q)$ a random variable distributed according to a negative binomial with parameters r and q, and $\mathrm{P}(\lambda)$ a random variable distributed according to a Poisson distribution with parameter λ.

1. The mean and variance of $\mathrm{NB}(r, q)$ are, respectively, $\lambda = \mathbf{E}(\mathrm{NB}(r, q)) = (1-q)r/q$ and $v = \mathbb{V}\mathrm{ar}(\mathrm{NB}(r, q)) = (1-q)r/q^2$. This represents an advantage over the use of a Poisson distribution where these two quantities can't be tuned independently one another, since $\mathbb{E}(\mathrm{P}(\lambda)) = \mathbb{V}\mathrm{ar}(\mathrm{P}(\lambda)) = \lambda$. Thus, the use of a negative binomial prior guarantees more flexibility.

2. The variance of the negative binomial distribution can be written in terms of the mean, according to the following formula: $v = \lambda + \frac{\lambda^2}{r}$. Hence the expert can choose the parameter λ equal to the number of distinct Y-STR haplotypes in nature, according to his expectation, and v equal to his opinion on the precision of this expectation.

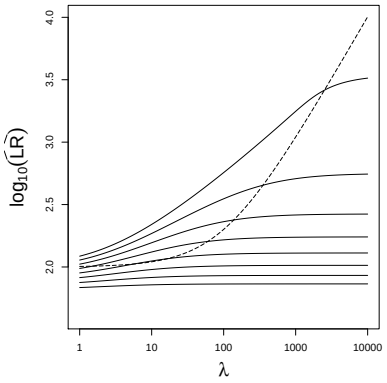
3. The negative binomial $\mathrm{NB}(r, q)$ is a Gamma mixture of Poisson.

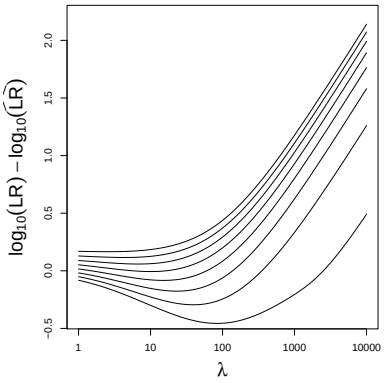
4. For fixed λ, when r increases, the negative binomial $\mathrm{NB}(r, q)$ tends in distribution to $\mathrm{P}(\lambda)$. This means that the negative binomial distribution can be seen as an extension of the Poisson distribution.

The same properties apply to the $[0, m]$-truncated case, both for the Negative Binomial, and for the Poisson, if m is big enough and the probability of 0 is small.

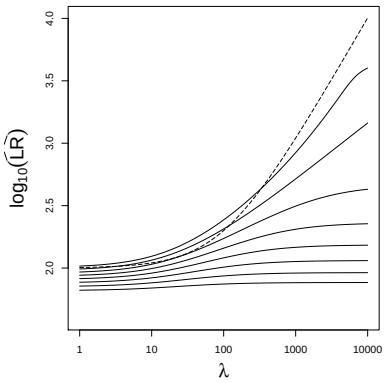
(a) (b)

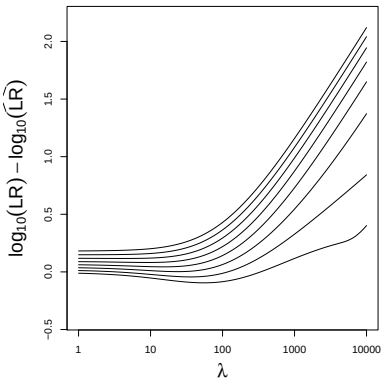
$r = 1$



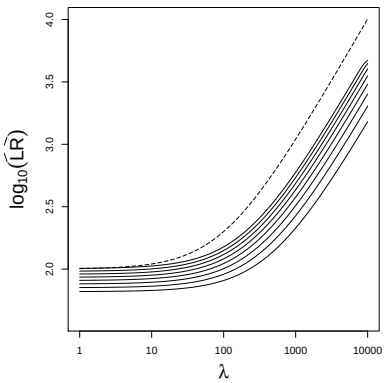


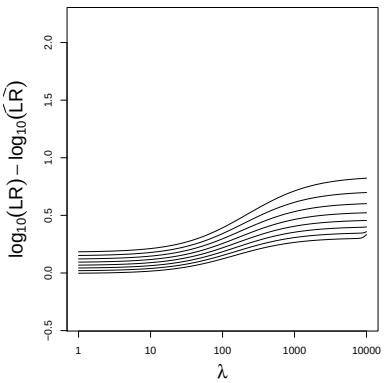
$r = 10$





$r = 100$





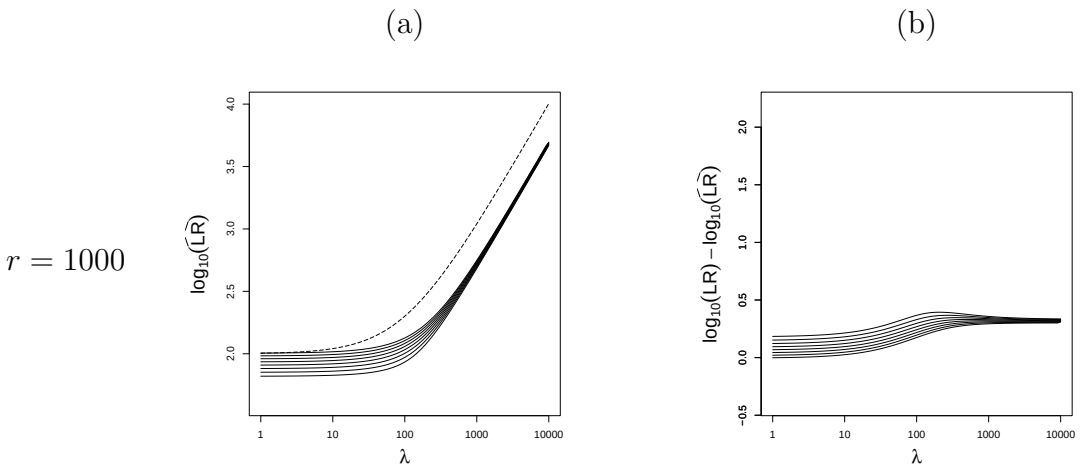


Figure 6.8: Sensitivity analysis for the three quantities $\log_{10} \mathrm{LR}$ (first column, black lines), $\log_{10} \widehat{\mathrm{LR}}$ (first column, dashed line) and the difference $\log_{10} \widehat{\mathrm{LR}} - \log_{10} \mathrm{LR}$ (second column) to different values of $\lambda = \mathbb{E}(K)$ (x axis) and of $k_{obs} \in \{30, 40, 50, 60, 70, 80, 90, 100\}$ (represented by the different lines, where the highest line corresponds to the highest k_{obs}).

6.7.3 Sensitivity analysis

The value of $\log_{10} \mathrm{LR}$ depends on the parameters of the prior (λ, r, and α) and on the values N and k_{obs}, regarding the data. Figure 6.8 represents the sensitivity of results to parameters λ and r in the rare type match case ($b_{e_s} = 0$) for fixed $\alpha = 1$, $N = 100$. It can be inferred from this analysis that when r increases the values depend more and more on λ and less and less on k_{obs}, and that for big r we fall back in the Poisson case, as explained in property 3.

According to the second column of Figure 6.8, one can see that for $r \geq 100$ the plug-in estimate always exceeds $\log_{10} \mathrm{LR}$. Anyway, the difference is important only if r is small, in particular for high values of λ.

6.7.4 Remarks about conventional priors

As mentioned above, the beta distribution and the Dirichlet distribution are the conventional choices for the prior on the parameter, in case of binomial or multinomial model, respectively. As stated in Curran et al. (2002), this "remains the accepted standard in some laboratories" because of the "appeal of simplicity and ease of implementation". Although we agree that this may have been a very sensible reason some decades ago, nowadays, with the computational skills provided by our computers, there are no more excuses to limit ourselves to these convenient priors. Indeed, a prior should reflect the expert's beliefs rather than standards of computational ease.

For the beta prior, the dependency of the LR results on the value of the hyperparameter α stresses once more the need of a different choice. Moreover, the model is profile-specific,

meaning that the beta priors is supposed to model the frequency of the profile observed at the crime scene, so that for a different scenario the model has to be changed.

For the Dirichlet prior, we have a similar issue. The dimension k of this prior should correspond to the number of different DNA types in the population. This problem, which for autosomal markers could be easily overcome, is more important when Y-STR haplotypes are considered, the state space being huge, and the database hardly representative. If we choose as k the number of different haplotypes observed in the database, then we are in trouble every time a new haplotype is observed, as for the situation of interest for this paper. By treating k as a Bayesian would do for an unknown quantity, we expected the likelihood ratio to depend a lot on the mean value of the prior chosen for k. The LR obtained with the Dirichlet method with all parameters $\alpha = 1$ turned out to depend only on the number of observed haplotypes in the database (and not on their frequencies). This is actually unattractive for Y-STR data, and is due to the symmetry: the data does not overrule the prior which makes *all* the positive p_i the same in size, and it is also the reason why the likelihood ratios obtained using the two methods (beta-binomial, and Dirichlet-multinomial) do not differ too much. Notice that for this prior we only focused on the case in which all the parameters α are equal to 1. More could have been done, for instance to explore the sensitivity of the likelihood ratio to changes in the α (Triggs and Curran, 2006), or to use hierarchical models (Chen and McGee, 2008). However, we preferred to investigate other types of prior (Cereda, 2016c) which we believe are more appropriate for Y-STR haplotypes frequencies. The two methods of Section 6.6 and Section 6.7 differ in the choice of the information retained from the database. The Beta binomial method only retains as information the relative frequency of the observed haplotype. A lot of information regarding other haplotypes is discarded, such as how many have been observed, and their frequencies. Let us point out that if there will ever be guidelines on how to choose the hyperparameters of the beta prior and of the Dirichlet prior, they should be compatible, meaning that the beta prior should be the one obtained from the Dirichlet by marginalisation.

6.8 Conclusion

This paper is intended to have several take-home messages. The first one is that a forensic statistician, before starting any evaluation, should make up his mind if he wants to use frequentist or Bayesian methods, since we have seen that the corresponding likelihood ratios are differently defined. If a Bayesian approach is chosen, which has the advantage that everything is combined into a single number, without any uncertainty involved, the LR should be calculated in a principled way. Bayesian plug-in (and frequentist plug-in), often proposed as proper Bayesian approach, can sometimes be seen as a convenient approximation of the Bayesian LR. However, the Bayesian plug-in is almost always anti-conservative in a way that is unfair to defence, and there are sometimes important differences with the full Bayesian method for particular choices of the hyperparameters of the priors. All this has been shown when the conventional choices for the prior (beta or Dirichlet) are made. The choice of the prior is an issue indeed. We believe that a true Bayesian should not make use of conventional priors, but of his own priors. Especially because, as shown, conventional choices lead to likelihood ratios which strongly depend on the hyperparameters of these priors. Choosing

more realistic prior may increase the difficulty of the computation of the likelihood ratio, but, also thanks to modern computational tools, this should not stop people from preferring them.

Acknowledgements

I am indebted to Charles Brenner for the useful discussion about this paper, which lead to many improvements. This research was supported by the Swiss National Science Foundation, through grants no. 105311-1445570 and 10531A-156146/1, and carried out in the context of a joint research project, supervised by Franco Taroni (University of Lausanne, Ecole des sciences criminelles), and Richard Gill (Mathematical Institute, Leiden University).

Chapter 7

Nonparametric Bayesian approach to LR assessment in case of rare type match

This chapter is based on:
Cereda, G. Nonparametric Bayesian approach to LR assessment in case of rare type match. *arXiv:1506.08444*. Submitted to: *Annals of Applied Statistics*.

Abstract

The evaluation of a match between the DNA profile of a stain found on a crime scene and that of a suspect (previously identified) involves the use of the unknown parameter $\mathbf{p} = (p_1, p_2, ...)$, (the ordered vector which represents the frequencies of the different DNA profiles in the population of potential donors) and the names of the different DNA profiles. We propose a Bayesian nonparametric method which models $\mathbf{p}$ through a random variable $\mathbf{P}$ distributed according to the two-parameter Poisson Dirichlet distribution, and discards the information about the names of the different DNA profiles. The ultimate goal of this model is to evaluate the so-called 'probative value' of DNA matches in the rare type case, that is the situation in which the suspect's profile, matching the crime stain profile, is not in the database of reference.

7.1 Introduction

The largely accepted method for evaluating how much some available data $\mathcal{D}$ (typically forensic evidence) is helpful in discriminating between two hypotheses of interest (the prosecution hypothesis H_p and the defense hypothesis H_d), is the calculation of the *likelihood ratio* (LR), a statistic that expresses the relative plausibility of the data under these hypotheses, defined as

$$\mathrm{LR} = \frac{\Pr(\mathcal{D}|H_p)}{\Pr(\mathcal{D}|H_d)}. \tag{7.1}$$

Widely considered the most appropriate framework to report a measure of the 'probative value' of the evidence regarding the two hypotheses (Robertson and Vignaux, 1995; Evett and Weir, 1998; Aitken and Taroni, 2004; Balding, 2005), it indicates the extent to which data is in favor of one hypothesis over the other. Forensic literature presents many approaches to calculate the LR, mostly divided into Bayesian and frequentist methods (see Cereda (2016a,b) for a careful differentiation between these two approaches).

This paper proposes a Bayesian nonparametric method for the LR assessment in the rare type match case, the challenging situation in which there is a match between some characteristic of the recovered material and of the control material, but this characteristic has not been observed before in previously collected samples (i.e. database of reference). This constitutes a problem because the value of the likelihood ratio depends on the unknown proportion of the matching characteristic in a reference population, and the uncertainty over this proportion is, in standard practice, dealt with using the relative frequency of the characteristic in the available database. In particular, we will focus on Y-STR data, for which the rare type match problem is often recurring (Cereda, 2016b).

To use a Bayesian nonparametric method we assume that there are infinitely many Y-STR profiles: the parameter of the model is the infinite dimensional vector $\mathbf{p}$, made of the (unknown) sorted population proportions of all possible Y-STR profiles. As prior over $\mathbf{p}$ we choose the two-parameter Poisson Dirichlet distribution, and we model the uncertainty over its own parameters α and θ through the use of an hyperprior. The information contained in the names of the profiles is discarded: this means to reduce the data $\mathcal{D}$ to a smaller amount of information D.

The paper is structured in the following way: Section 7.2 introduces the notation, the assumptions of our model and the prior distribution chosen for parameter $\mathbf{p}$. Section 7.3 presents the model, along with some theory on random partitions useful to provide a convenient and compact representation of the reduced data D. An alternative representation of the same model via the two-parameter Chinese restaurant process is also described. Section 7.4 introduces relevant known results regarding the two-parameter Poisson Dirichlet distribution, along with a new lemma that will allow to derive the likelihood ratio in a very elegant way (Section 7.5).

Section 7.6 proposes the application of this model to a real database of Y-STR profiles. We will discuss data driven choices for the hyperpriors, and comparison with the frequentist likelihood ratio values obtained both reducing and not the data in the ideal situation in which vector $\mathbf{p}$ is known.

7.2 A Bayesian nonparametric model for the rare type match

7.2.1 The rare type match problem

The evaluation of a match between the profile of a particular piece of evidence and a suspect's profile depends on the proportion of that profile in the population of potential perpetrators. Indeed, it is intuitive that the rarer the matching profile, the more the suspect is in trouble.

Problems arise when the observed frequency of the profile in a sample from the population of interest (i.e., in a reference database) is 0. Such characteristic is likely to be rare, but it is challenging to quantify how rare it is. The rare type match problem is particularly important in case a new kind of forensic evidence, such as results from DIP-STR markers (see for instance Cereda et al. (2014a)) is involved, and for which the available database size is still limited. The same happens when Y-chromosome (or mitochondrial) DNA profiles are used, since the set of possible Y-STR profiles is extremely large. As a consequence, most of the Y-STR haplotypes are not represented in the database. The Y-STR marker system will thus be retained here as an extreme but in practice common and important way in which the problem of assessing evidential value of rare type match can arise. This problem is so substantial that it has been defined "the fundamental problem of forensic mathematics" (Brenner, 2010).

The *empirical frequency estimator*, also called *naive estimator*, that uses the frequency of the characteristic in the database, puts unit probability mass on the set of already observed characteristics, and it is thus unprepared for the observation of a new type. A solution could be the *add-constant* estimators (in particular the well-known *add-one* estimator, due to Laplace (1814), and the *add-half* estimator of Krichevsky and Trofimov (1981)), which add a constant to the count of each type, included the unseen ones. However, this method requires to know the number of possible unseen types, and it performs badly when this number is large compared to the sample size (see Gale and Church (1994) for an additional discussion). Alternatively, Good (1953), based on an intuition on A.M. Turing, proposed the *Good Turing estimator* for the total unobserved probability mass, based on the proportion of singleton observations in the sample. An extension of this estimator is applied to the frequentist LR assessment in the rare type match case in Cereda (2016b). For a comparison between *add one* and *Good-Turing* estimator, see Orlitsky et al. (2003). As pointed out in Anevski et al. (2013), the *naive estimator*, and the *Good Turing estimator* are in some sense complementary: the first gives a good estimate for the observed types, and the second for the probability mass of the unobserved ones. More recently, Orlitsky et al. (2004) have introduced the *high profile estimator*, which extends the tail of the *naive estimator* to the region of unobserved types. Anevski et al. (2013) improved this estimator and provided the consistency proof. Papers that address the rare Y-STR haplotype problem in forensic context are for instance Egeland and Salas (2008), Brenner (2010), and Cereda (2016a). The latter applies the classical Bayesian approach (the beta binomial and the Dirichlet multinomial problem) to the LR assessment in the rare haplotype case. Moreover, the Discrete Laplace method presented in Andersen et al. (2013b), even though not specifically designed for the rare type match case, can be successfully applied to that purpose (Cereda, 2016b).

Bayesian nonparametric estimators for the probability of observing a new type have been proposed by Tiwari and Tripathi (1989) using Dirichlet process, by Lijoi et al. (2007) using general Gibbs prior, and by Favaro et al. (2009) with specific interest to the two-parameter Poisson Dirichlet prior. However, the LR assessment requires not only the probability of observing a new species but also the probability of observing this same species twice (according to the defense the crime stain profile and the suspect profile are two independent observations): to our knowledge, the present paper is the first one to address the problem of LR assessment in the rare haplotype case using Bayesian nonparametric models. As prior for $\mathbf{p}$ we will use the two-parameter Poisson Dirichlet distribution, which is proving useful in many

discrete domains, in particular language modelling (Teh et al., 2006). In addition, it shows a power-law behaviour which describes an incredible variety of phenomena (Newman, 2005). Indeed it can be proved that

7.2.2 Notation

Throughout the paper the following notation is chosen: random variables and their values are denoted, respectively, with uppercase and lowercase characters: x is a realization of X. Random vectors and their values are denoted, respectively, by uppercase and lowercase bold characters: $\mathbf{p}$ is a realization of the random vector $\mathbf{P}$. Probability is denoted with $\Pr(\cdot)$, while density of a continuous random variable X is denoted alternatively by $p_X(x)$ or by $p(x)$ when the subset is clear from the context. For a discrete random variable Y, the density notation $p_Y(y)$ and the discrete one $\Pr(Y = y)$ will be alternately used. Moreover, we will use shorthand notation like $p(y \mid x)$ to stand for the probability density of Y with respect to the conditional distribution of Y given $X = x$.

Notice that in Formula (7.1), $\mathcal{D}$ was regarded as the event corresponding to the observation of the available data. However, later in the paper, $\mathcal{D}$ will be regarded as a random variable generically representing the data. The particular data at hand will correspond to the value d. In that case, the following notation will thus be preferred:

$$\mathrm{LR} = \frac{\Pr(\mathcal{D} = d|H = h_p)}{\Pr(\mathcal{D} = d|H = h_d)} \quad \text{or} \quad \frac{p(d|h_p)}{p(d|h_d)}. \tag{7.2}$$

Lastly, notice that "DNA types" is used throughout the paper as a general term to indicate Y-STR profiles.

7.2.3 Model assumptions

Our model is based on the two following assumptions:

Assumption 1 There are infinitely many DNA types in Nature.

The reason for this assumption is that there are so many possible DNA types that they can be considered infinite. This assumption, already used by e.g. Kimura (1964) in the 'infinite alleles model', allows to use Bayesian nonparametric methods and avoids the problem of specifying how many different types there are in Nature.

Assumption 2 The names of the different DNA types do not contain information.

Actually, the specific sequence of numbers that forms a DNA profile carries information: if two profiles show few differences this means that they are separated by few mutation drifts, hence the profiles share a relatively recent common ancestor. However, this information is difficult to exploit and may be not so relevant for the LR assessment. This is the reason why we will treat DNA types as "colors", and only consider the partition into different categories. Stated otherwise, we put no topological structure on the space of the DNA types.

Notice that this assumption makes the model a priori suitable for any characteristic which shows many different possible types, thus what written still holds, in principle, also replacing 'DNA types' with any other category. However, in this paper we will only test the model with Y-STR profiles as categories.

7.2.4 Prior

In Bayesian statistics, parameter of interest are modeled through random variables. The (prior) distribution over a parameter should represent the uncertainty about its value.

LR assessment for the rare type match involves two unknown parameters of interest: one is $h \in \{h_p, h_d\}$, representing the unknown true hypothesis, the other is $\mathbf{p}$, the vector of the unknown population frequencies of all DNA profiles in the population of potential perpetrators. The dichotomous random variable H is used to model parameter h, and the posterior distribution of this random variable, given the data, is the ultimate aim of the forensic inquiry. In a similar way, random variable $\mathbf{P}$ is used to model the uncertainty over $\mathbf{p}$. Because of Assumption 1, $\mathbf{p}$ is an infinite dimensional parameter, hence the need of Bayesian nonparametric methods (Hjort et al., 2010). In particular, $\mathbf{p} = (p_t | t \in T)$, with T a countable set of indexes, $p_t > 0$, and $\sum_t p_t = 1$. Moreover, because of Assumption 2, data can be reduced to random partitions, as explained in Section 7.3.1, and it will turn out that the distribution of these partitions does not depend on the order of the p_i. Hence, we can force the parameter $\mathbf{p}$ to have values in $\nabla_\infty = \{(p_1, p_2, ...) | p_1 \geq p_2 \geq ..., \sum p_i = 1, p_i > 0\}$, the ordered infinite dimensional simplex. The uncertainty about its value is expressed by the prior distribution over $\mathbf{p}$, for which we choose the two-parameter Poisson Dirichlet distribution (Pitman and Yor, 1997; Feng, 2010; Buntine and Hutter, 2010; Carlton, 1999; Pitman and Picard, 2006), defined in the following way:

Definition 1 (two-parameter GEM distribution)**.** *Given α and θ satisfying the following conditions:*

$$0 \leq \alpha < 1, \text{ and } \theta > -\alpha. \tag{7.3}$$

the vector $\mathbf{W} = (W_1, W_2, ...)$ is said to be distributed according to the GEM(α, θ)*, if*

$$\forall i \quad W_i = V_i \prod_{j=1}^{i-1} (1 - V_j),$$

where V_1, V_2,... are independent random variables distributed according to

$$V_i \sim B(1 - \alpha, \theta + i\alpha).$$

It holds that $W_i > 0$, and $\sum_i W_i = 1$.

The GEM distribution (short for Griffin - Engen - McCloskey distribution') is well known in literature as the "stick breaking prior", since it measures the random sizes in which a stick is broken iteratively. This distribution is invariant under size-biased permutations (Engen, 1975), that is the random permutation defined by sampling from the population and assigning to each type a label, based on the order in which the types are first sampled.

Definition 2 (Two-parameter Poisson Dirichlet distribution)**.** *Given α and θ satisfying condition* (7.3)*, and a vector $\mathbf{W} = (W_1, W_2, ...) \sim GEM(\alpha, \theta)$, the random vector $\mathbf{P} = (P_1, P_2, ...)$ obtained by ordering $\mathbf{W}$, such that $P_i \geq P_{i+1}$, is said to be* Poisson Dirichlet distributed *$PD(\alpha, \theta)$. Parameter α is called* discount parameter*, while θ is the* concentration parameter.

Notice that the vector $\mathbf{P}$ is obtained by sorting the vector $\mathbf{W}$ in nonincreasing order, while the vector $\mathbf{W}$ can be obtained (in distribution) by the so-called *size-biased permutation* of the indexes of $\mathbf{P}$ (Perman et al., 1992; Pitman and Yor, 1997).

The two-parameter Poisson Dirichlet distribution $\text{PD}(\alpha, \theta)$ is the generalization of the well-known Poisson Dirichlet distribution with a single parameter θ introduced by Kingman (1975), which is the representation measure (Kingman, 1977, 1978) of the celebrated *Ewens sampling formula* (Ewens, 1972), widely applied in genetics (Karlin and McGregor, 1972; Kingman, 1980). For our model we will not allow $\alpha = 0$, hence we will assume $0 < \alpha < 1$.

It is worth mentioning that an alternative choice for the parameters space is $\alpha < 0$, $\theta = -m\alpha$ for some $m \in \mathbb{N}$ (Pitman, 1996; Gnedin and Pitman, 2006; Gnedin, 2009; Cerquetti, 2010). It corresponds to a model with finitely many (m) DNA types, where $\mathbf{P} = (P_1, ..., P_m)$ is Dirichlet distributed with m parameters equal to $-\alpha$. We will not consider this case.

Lastly, we point out that, in practice, we cannot assume to know parameters α and θ: we will model the uncertainty about them using an hyperprior.

7.3 The model

The typical data to evaluate in case of a match is $\mathcal{D} = (E, B)$, where $E = (E_s, E_t)$, and

- E_s = suspect's DNA type,
- E_t = crime stain's DNA type (matching with the suspect's type),
- B = a reference database of size n, which contains a sample of DNA types, indexed by $i = 1, ..., n$, from the population of possible perpetrators.

The hypotheses of interest for the case are:

- h_p = The crime stain was left by the suspect,
- h_d = The crime stain was left by someone else.

In agreement with Assumption 2, the model will ignore information about the names of the DNA types: data $\mathcal{D} = (E, B)$ will be reduced to D accordingly. The Bayesian network of Figure 7.1 encapsulates the conditional dependencies of the random variables of the proposed model:

- H is a dichotomous random variable that represents the hypotheses of interest and can take values $h \in \{h_p, h_d\}$, according to the prosecution or the defense, respectively. A uniform prior on the hypotheses is chosen:

$$\Pr(H = h) \propto 1 \quad \text{for } h \in \{h_p, h_d\}.$$

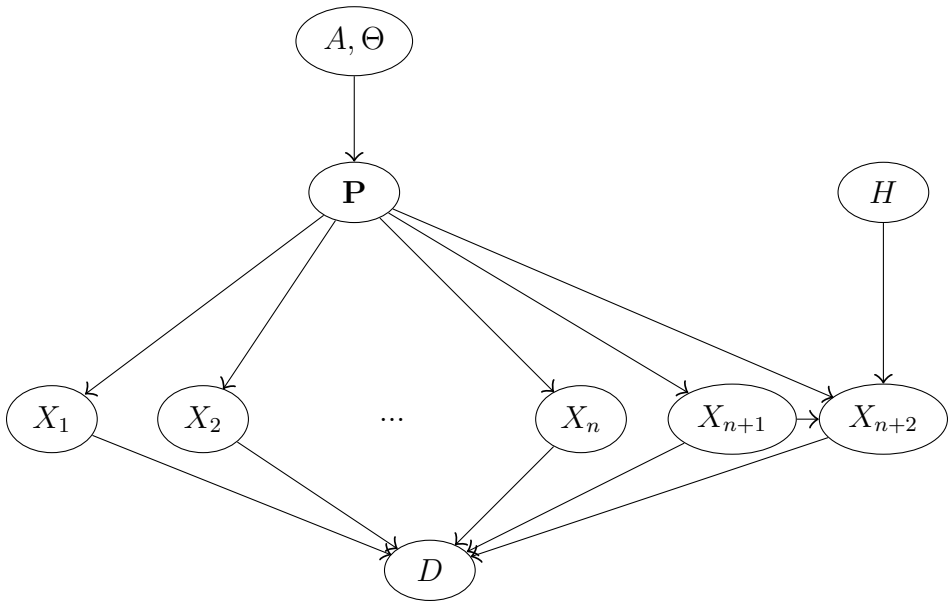


Figure 7.1: Bayesian network to show the conditional dependencies of the relevant random variables in our model.

Notice that this choice is made for mathematical convenience, since it will not affect the likelihood ratio.

- (A, Θ) is the random vector that represents the hyperparameters α and θ, satisfying condition (7.3). The joint prior density of these two parameters (hyperprior) will be generically denoted as $p(\alpha, \theta)$:

$$(A, \Theta) \sim p(\alpha, \theta).$$

- The random vector $\mathbf{P}$ with values in ∇_∞, represents the ranked population frequencies. $\mathbf{P} = (p_1, p_2, ...)$ means that p_1 is the frequency of the most common DNA type in the population, p_2 is the frequency of the second most common DNA type, and so on. As a prior for $\mathbf{P}$ we use the two-parameter Poisson Dirichlet distribution (see Definition 2):

$$\mathbf{P}|A = \alpha, \Theta = \theta \sim PD(\alpha, \theta).$$

- The database is assumed to be a random sample from the population. Integer valued random variables X_1, ..., X_n are here used to represent the ranks of the population proportions of the DNA types in the database. For instance, $X_3 = 5$ means that the third individual in the database has the fifth most common DNA type in the population. Given $\mathbf{p}$ they are an i.i.d. sample from $\mathbf{p}$:

$$X_1, X_2, ..., X_n|\mathbf{P} = \mathbf{p} \sim_{i.i.d.} \mathbf{p}. \tag{7.4}$$

  To observe X_1, ..., X_n, one would need to know the rank, in terms of population proportion, of the frequency of each DNA types in the database. This is not known, hence we don't observe $X_1, ..., X_n$.

- X_{n+1} represents the rank of the suspect's DNA type. It is again an indipendent draw from $\mathbf{p}$.

$$X_{n+1}|\mathbf{P} = \mathbf{p} \sim \mathbf{p}.$$

- X_{n+2} represents the rank of the crime stain's DNA type. According to the prosecution, given $X_{n+1} = x_{n+1}$, this random variable is deterministic (it is equal to x_{n+1} with probability 1). According to the defense it is another sample from $\mathbf{p}$, independent of the previous ones:

$$X_{n+2}|\mathbf{P} = \mathbf{p}, X_{n+1} = x_{n+1}, H = h \sim \begin{cases} \delta_{x_{n+1}} & \text{if } h = h_p \\ \mathbf{p} & \text{if } h = h_d \end{cases}.$$

As already mentioned, $X_1, ..., X_{n+2}$ cannot be observed. They represent the database, where the names of the DNA types have been replaced by their (unknown) ranks in $\mathbf{p}$, and constitute an intermediate layer.

Section 7.3.1 recalls some notions about random partitions, useful before defining node D, the 'reduced' data that we want to evaluate.

7.3.1 Random partitions

A *partition of a set* A is an unordered collection of nonempty and disjoint subsets of A the union of which forms A. Particularly interesting for our model are partitions of the set $A = [n] = \{1, ..., n\}$, denoted as $\pi_{[n]}$. The set of all partitions of $[n]$ will be denoted as $\mathcal{P}_{[n]}$. Random partitions of $[n]$ will be denoted as $\Pi_{[n]}$. In addition, a *partition of* n is a finite nonincreasing sequence of positive integers that sum up to n. Partitions of n will be denoted as π_n, random partitions as Π_n.

Given a sequence of integer valued random variables $X_1, ..., X_n$, let $\Pi_{[n]}(X_1, X_2, ..., X_n)$ be the random partition defined by the equivalence classes of their indexes using the random equivalence relation $i \sim j$ if and only if $X_i = X_j$. This construction allows to build a map from the set of values of $X_1, ..., X_n$ to the set of the partitions of $[n]$ as in the following example ($n = 10$):

$$\begin{aligned} \mathbb{N}^{10} &\to \mathcal{P}_{[10]} \\ X_1, ..., X_{10} &\longmapsto \Pi_{[10]}(X_1, X_2, ..., X_{10}) \\ (2, 4, 2, 4, 3, 3, 10, 13, 5, 4) &\longmapsto \{\{1, 3\}, \{2, 4, 10\}, \{5, 6\}, \{7\}, \{8\}, \{9\}\} \end{aligned}$$

In agreement with Assumption 2, in our model we can consider the reduction of data which ignores information about the names of the DNA types: this is achieved, for instance, by retaining from the database only the equivalence classes of the indexes of the individuals, according to the equivalence relation "to have the same DNA type". Stated otherwise, the database is reduced to the partition $\pi_{[n]}^{\mathrm{Db}}$, obtained using these equivalence classes. However, data is not only made of the database B. There are also two new DNA profiles which are equal one another and different from the already observed ones. When the suspect's profile is considered we obtain the partition $\pi_{[n+1]}^{\mathrm{Db}+}$, where the first n integers are partitioned as in

$\pi^{\text{Db}}_{[n]}$, and $n+1$ constitutes a class by itself (at least in the rare type match case). When the crime stain profile is considered we obtain the partition $\pi^{\text{Db++}}_{[n+2]}$ where the first n integers are partitioned as in $\pi^{\text{Db}}_{[n]}$, and $n+1$ and $n+2$ belongs to the same (new) class.

Random variables $\Pi^{\text{Db}}_{[n]}$, $\Pi^{\text{Db+}}_{[n+1]}$, and $\Pi^{\text{Db++}}_{[n+2]}$ are used to model $\pi^{\text{Db}}_{[n]}$, $\pi^{\text{Db+}}_{[n+1]}$, and $\pi^{\text{Db++}}_{[n+2]}$, respectively.

Since prosecution and defense agree on the distribution of $X_1, ..., X_{n+1}$, but not on the distribution of X_{n+2}, they also agree on the distribution of $\Pi^{\text{Db+}}_{[n+1]}$ but disagree on the distribution of $\Pi^{\text{Db++}}_{[n+2]}$.

The crucial point of the model is that, by construction, the same random partitions can be defined through random variables X_1, ..., X_{n+2}. Indeed, it holds that:

$$\begin{aligned}\Pi^{\text{Db}}_{[n]} &= \Pi_{[n]}(X_1, ..., X_n),\\ \Pi^{\text{Db+}}_{[n+1]} &= \Pi_{[n+1]}(X_1, ..., X_{n+1}),\\ \Pi^{\text{Db++}}_{[n+2]} &= \Pi_{[n+2]}(X_1, ..., X_{n+2}).\end{aligned}$$

Moreover, although X_1, ..., X_{n+2} were not observable, the random partitions $\Pi^{\text{Db}}_{[n]}, \Pi^{\text{Db+}}_{[n+1]}$, and $\Pi^{\text{Db++}}_{[n+2]}$ are observable.

To clarify, consider the following example of a database (B) with $k = 6$ different DNA types, from $n = 10$ individuals:

$$B = (h_1, h_2, h_1, h_2, h_3, h_3, h_4, h_5, h_6, h_2),$$

where h_i is the name of the ith DNA type according to the order chosen for the database. This can be reduced to the partition of [10]:

$$\pi^{\text{Db}}_{[10]} = \{\{1, 3\}, \{2, 4, 10\}, \{5, 6\}, \{7\}, \{8\}, \{9\}\}.$$

Then, the part of data whose distribution is agreed on by prosecution and defense is

$$\pi^{\text{Db+}}_{[11]} = \{\{1, 3\}, \{2, 4, 10\}, \{5, 6\}, \{7\}, \{8\}, \{9\}, \{11\}\},$$

while the entire (reduced) data D can be represented as

$$\pi^{\text{Db++}}_{[12]} = \{\{1, 3\}, \{2, 4, 10\}, \{5, 6\}, \{7\}, \{8\}, \{9\}, \{11, 12\}\}.$$

Now, assume that we know the rank in the population of each of the DNA types in the database: we know that h_1 is, for instance, the second most frequent type, h_2 is the fourth most frequent type, and so on. Stated otherwise, we are now assuming that we observe the variables X_1, ..., X_{n+2}: for instance, $X_1 = 2$, $X_2 = 4$, $X_3 = 2$, $X_4 = 4$, $X_5 = 3$, $X_6 = 3$, $X_7 = 10$, $X_8 = 13$, $X_9 = 5$, $X_{10} = 4$, $X_{11} = 9$, $X_{12} = 9$. It is easy to check that $\Pi_{[10]}(X_1, ..., X_{10}) = \pi^{\text{Db}}_{[10]}$, $\Pi_{[11]}(X_1, ..., X_{11}) = \pi^{\text{Db+}}_{[11]}$, and $\Pi_{[12]}(X_1, ..., X_{12}) = \pi^{\text{Db++}}_{[12]}$.

As already mentioned, data D is defined as:

- $D = \pi^{\text{Db++}}_{[n+2]}$, obtained partitioning the database enlarged with the two new observations (or partitioning $X_1, ..., X_{n+2}$).

Node D of Figure 7.1 is defined accordingly. Notice that, given $X_1, ..., X_{n+2}$, D is deterministic. An important result is that, according to Proposition 4 in Pitman (1992) it is possible to derive directly the distribution of $D \mid \alpha, \theta, H$. In particular, it holds that if

$$\mathbf{P} \mid \alpha, \theta \sim PD(\alpha, \theta),$$

and

$$X_1, X_2, ... \mid \mathbf{P} = \mathbf{p} \sim_{\text{i.i.d}} \mathbf{p},$$

then, for all $n \in \mathbb{N}$, the random partition $\Pi_{[n]} = \Pi_{[n]}(X_1, ..., X_n)$ has the following distribution:

$$\mathbb{P}^{\alpha,\theta}_n(\pi_{[n]}) := \Pr(\Pi_{[n]} = \pi_{[n]} | \alpha, \theta) = \frac{[\theta + \alpha]_{k-1;\alpha}}{[\theta + 1]_{n-1;1}} \prod_{i=1}^{k} [1 - \alpha]_{n_i - 1;1}, \tag{7.5}$$

where n_i is the size of the ith block of $\pi_{[n]}$ (the blocks are here ordered according to the least element), and $\forall x, b \in \mathbb{R}, a \in \mathbb{N}$, $[x]_{a,b} := \begin{cases} \prod_{i=1}^{a-1}(x + ib) & \text{if } a \in \mathbb{N}\backslash\{0\} \\ 1 & \text{if } a = 0 \end{cases}$. This formula is also known as the *Pitman sampling formula*, further studied in Pitman (1995). Notice that for $\alpha = 0$ we obtain the Ewens's sampling formula.

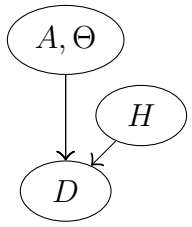


Figure 7.2: Simplified version of the Bayesian network in Figure 7.1

It follows that we can get rid of the intermediate layer of nodes X_1, ..., X_{n+2}, and $\Pr(D|\alpha, \theta, h_p) = \mathbb{P}^{\alpha,\theta}_{n+1}(\pi^{\text{Db+}}_{[n+1]})$, while $\Pr(D|\alpha, \theta, h_d) = \mathbb{P}^{\alpha,\theta}_{n+2}(\pi^{\text{Db++}}_{[n+2]})$. The model of Figure 7.1 can thus be simplified to the one in Figure 7.2.

7.3.2 Chinese Restaurant representation

There is an alternative characterization of this model, called "Chinese restaurant process", due to Aldous (1985) for the one parameter case, and studied in details for the two-parameter version in Pitman and Picard (2006). It is defined as follows: consider a restaurant with infinite many tables, each one infinitely large. Let $Y_1, Y_2, ...$ be integer valued random variables that represent the seating plan: tables are ranked in order of occupancy, and $Y_i = j$ means that the ith customer seats at the jth table to be created. The process is described by the following transition matrix:

$$Y_1 = 1,$$

$$\Pr(Y_{n+1} = i | Y_1, ..., Y_n) = \begin{cases} \dfrac{\theta + k\alpha}{n + \theta} & \text{if } i = k + 1 \\ \\ \dfrac{n_i - \alpha}{n + \theta} & \text{if } 1 \leq i \leq k \end{cases} \tag{7.6}$$

where k is the number of tables occupied by the first n customers, and n_i is the number of customers that occupy table i. The process depends on two parameters α and θ with the same conditions (7.3).

$Y_1, ..., Y_n$ are not i.i.d., nor exchangeable, but it holds that $\Pi_{[n]}(Y_1, ..., Y_n)$ is distributed as $\Pi_{[n]}(X_1, ..., X_n)$, with $X_1, ..., X_n$ defined as in (7.4) (in particular they are both distributed according to the Pitman sampling formula (7.5)).

Stated otherwise, we can use the seating plan of n customers $Y_1, ..., Y_n$, or $X_1, ..., X_n$ (the database) and we obtain the same partition $\pi^{\text{Db}}_{[n]}$. Similarly $\pi^{\text{Db+}}_{[n+1]}$ is obtained when a new customer has chosen an unoccupied table (remember we are in the rare type match case), and $\pi^{\text{Db++}}_{[n+2]}$ is obtained when the $n + 2$nd customer goes to the table already chosen by the $n + 1$st customer(suspect and crime stain have the same DNA type). In particular, thanks to (7.6), we can write

$$p(\pi^{\text{Db++}}_{[n+2]} \mid h_p, \pi^{\text{Db+}}_{[n+1]}, \alpha, \theta) = 1, \tag{7.7}$$

and

$$p(\pi^{\text{Db++}}_{[n+2]} \mid h_d, \pi^{\text{Db+}}_{[n+1]}, \alpha, \theta) = \frac{1 - \alpha}{n + 1 + \theta}, \tag{7.8}$$

since the $n+2$nd customer goes to the same table as the $n+1$st (who was sitting alone).

7.4 Some results

This section presents some useful results that will be used in the forthcoming sections. In particular, Lemma 2, suitable to broader applications, is here applied to simplify the likelihood ratio development. Then, some results from Pitman and Picard (2006) regarding the two-parameter Poisson Dirichlet distribution, are listed.

7.4.1 A useful Lemma

The following lemma is a result regarding four general random variables A, X, Y, H whose conditional dependencies are described by the Bayesian network of Figure 8.4. The importance of this result is due to the possibility of applying it to a very common forensic situation: the prosecution and the defense disagree on the distribution of the entirety of data (Y) but agree on the distribution of a part it (X), and these distributions depend on parameters (A).

Lemma 2. *Given four random variables A, H, X and Y, whose conditional dependencies are represented by the Bayesian network of Figure 7.3, the likelihood function for h, given $X = x$ and $Y = y$ satisfies*

$$\text{lik}(h \mid x, y) \ \propto \mathbb{E}(p(y \mid x, A, h) \mid X = x).$$

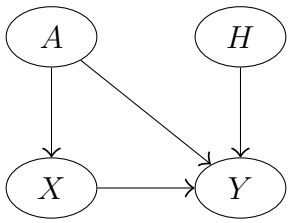
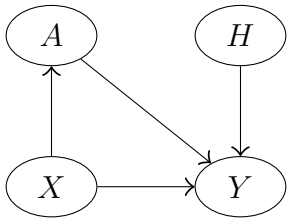


Figure 7.3: Conditional dependencies of the random variables of Lemma 2

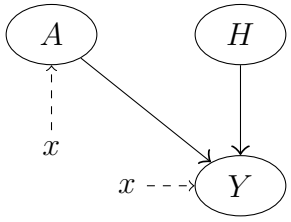
Proof. The model of Figure 7.3 represents four variables A, H, X and Y whose joint probabilty density can be factored as

$$p(a, h, x, y) \;=\; p(a)\, p(x \mid a)\, p(h)\, p(y \mid x, a, h).$$

By Bayes formula, $p(a)\, p(x \mid a) = p(x)\, p(a \mid x)$. This rewriting corresponds to reversing the direction of the arrow between A and X:



The random variable X is now a root node. This means that when we probabilistically condition on $X = x$, the graphical model changes in a simple way: we can delete the node X, but just insert the value x as a parameter in the conditional probability tables of the variables A and Y which formerly had an arrow from node X. The next graph represents this model:



This tells us, that conditional on $X = x$, the joint density of A, Y and H is equal to

$$p(a \mid x)p(h)p(y \mid x, a, h).$$

The joint density of H and Y is obtained by integrating out the variable a. It can be expressed as a conditional expectation value, since $p(a \mid x)$ is the density of A given $X = x$. We find:

$$p(h)\mathbb{E}(p(y \mid x, A, h) \mid X = x).$$

Recall that this is the joint density of two of our variables, H and Y, after conditioning on the value $X = x$. Let us now also condition on $Y = y$. It follows that the density of H given $X = x$ and $Y = y$ is proportional (as function of H, for fixed x and y) to the same expression, $p(h)\mathbb{E}(p(y \mid x, A, h) \mid X = x)$.

This is a product of the prior for h with some function of x and y. Since posterior odds equals prior odds times likelihood ratio, it follows that the likelihood function for h, given $X = x$ and $Y = y$ satisfies

$$\text{lik}(h \mid x, y) \ \propto \mathbb{E}(p(y \mid x, A, h) \mid X = x).$$

□

Corollary 3. *Given four random variables A, H, X and Y, whose conditional dependencies are represented by the network of Figure 8.4, the likelihood ratio for $H = h_1$ against $H = h_2$ given $X = x$ and $Y = y$ satisfies*

$$\text{LR} = \frac{\mathbb{E}(p(y|x, A, h_1)|X = x)}{\mathbb{E}(p(y|x, A, h_2)|X = x)}. \tag{7.9}$$

The importance of Lemma 2, and Corollary 3 is due to the possibility of applying it to our model. Indeed, as already noticed, since defense and prosecution agree on the distribution of $\pi^{\text{Db+}}_{[n+1]}$, but not on the distribution of $\pi^{Db++}_{[n+2]}$, and data depends on parameters α and θ.

7.4.2 Known results about the two-parameter Poisson Dirichlet distribution

We will now list some theoretical results which will be useful in the forthcoming analysis. Most of these results can be found in Pitman and Picard (2006).

Denote as K_n the random number of blocks of a partition $\Pi_{[n]}$ distributed according to the Pitman sampling formula with parameters α and θ.

- There exists a positive random variable S_α such that

$$\lim_{n\to+\infty} \frac{K_n}{n^\alpha} = S_\alpha \quad \text{a.s.} \tag{7.10}$$

  the distribution of S_α is a generalization of the Mittag-Leffler distribution (Gorenflo et al., 2014).

- If $\mathbf{P} \sim \text{PD}(\alpha, \theta)$, then

$$\frac{P_i}{Zi^{-1/\alpha}} \to 1, \quad \text{a.s., when } i \to +\infty \tag{7.11}$$

  for a random variable Z such that $Z^{-\alpha} = \Gamma(1-\alpha)/S_\alpha$.

- For a fixed $\alpha \in (0, 1)$, the $\text{PD}(\alpha, \theta)$ (for different θ) are all mutually absolutely continuous. This means that θ cannot be consistently estimated for α in the range of interest. On the other hand, the power-baw behavior described above tells us that α can be consistently estimated.

- Studying (7.6) one can see that when n increases, the parameter θ becomes less and less important. However, it describes how much "social" are the customers: the smaller θ the more the customers tend to seat to already occupied tables. Thus, it determines the sizes of the big tables, but it won't be much important for our application (the more rare DNA types correspond to small tables).
- Given Π_n distributed according to Pitman sampling formula (7.5), it holds that

$$\lim_{n\to+\infty} \frac{m_j(n)}{n^\alpha} = \frac{\alpha\Gamma(j-\alpha)}{\Gamma(1-\alpha)j!} S_\alpha \quad \text{a.s.} \, \forall j \tag{7.12}$$

  where $m_j(n)$, $j = 1, ..., n$ the random number of blocks of the partition $\Pi_{[n]}$ of size j. This result is presented in Gnedin et al. (2007), based on Karlin (1967).

7.5 The likelihood ratio

Using the hypotheses and the reduction of data D defined in Section 7.3, the likelihood ratio will be defined as

$$\text{LR} = \frac{p(\pi^{\text{Db++}}_{[n+2]}|h_p)}{p(\pi^{\text{Db++}}_{[n+2]}|h_d)} = \frac{p(\pi^{\text{Db+}}_{[n+1]}, \pi^{\text{Db++}}_{[n+2]}|h_p)}{p(\pi^{\text{Db+}}_{[n+1]}, \pi^{\text{Db++}}_{[n+2]}|h_d)}.$$

The last equality holds due to the fact that $\Pi^{\text{Db+}}_{[n+1]}$ is a deterministic function of $\Pi^{\text{Db++}}_{[n+2]}$.

Now, we can apply Corollary 3 with (A, Θ) playing the role of A, $X = \Pi^{\text{Db+}}_{[n+1]}$, and $Y = \Pi^{\text{Db++}}_{[n+2]}$ to obtain:

$$\begin{aligned}\text{LR} &= \frac{\mathbb{E}(p(\pi^{\text{Db++}}_{[n+2]} \mid \pi^{\text{Db+}}_{[n+1]}, A, \Theta, h_p) \mid \Pi^{\text{Db+}}_{[n+1]} = \pi^{\text{Db+}}_{[n+1]})}{\mathbb{E}(p(\pi^{\text{Db++}}_{[n+2]} \mid \pi^{\text{Db+}}_{[n+1]}, A, \Theta, h_d) \mid \Pi^{\text{Db+}}_{[n+1]} = \pi^{\text{Db+}}_{[n+1]})} \\ &= \frac{1}{\mathbb{E}\left(\frac{1-A}{n+1+\Theta} \mid \Pi^{\text{Db+}}_{[n+1]} = \pi^{\text{Db+}}_{[n+1]}\right)}.\end{aligned}$$

where the last equality is due to (7.7) and (7.8). By defining the random variable $\Phi = n\dfrac{1-A}{n+1+\Theta}$ we can write the LR as

$$\text{LR} = \frac{n}{\mathbb{E}(\Phi \mid \Pi^{\text{Db+}}_{[n+1]} = \pi^{\text{Db+}}_{[n+1]})}. \tag{7.13}$$

7.5.1 True LR

It is now interesting to study the frequentist likelihood ratio values obtained with (7.13), and to compare it with the 'true' ones, meaning the LR values obtained when vector $\mathbf{p}$ is known. This corresponds to having the list of the frequencies of all the DNA types in the population of interest. Then, the model can be represented by the Bayesian network of Figure 7.4.

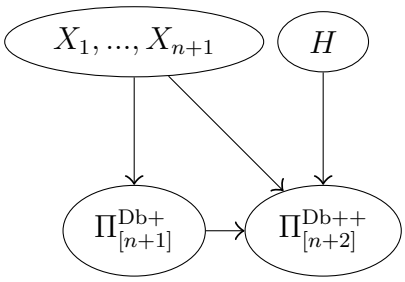


Figure 7.4: Bayesian network for the case in which $\mathbf{p}$ is known.

The LR in this case can be obtained using again Corollary 3, where now X_1, ..., X_{n+1} play the role of A.

$$\text{LR}_{|\mathbf{p}} = \frac{p(\pi^{\text{Db++}}_{[n+2]}, \pi^{\text{Db+}}_{[n+1]} \mid h_p, \mathbf{p})}{p(\pi^{\text{Db++}}_{[n+2]}, \pi^{\text{Db+}}_{[n+1]} \mid h_d, \mathbf{p})} \tag{7.14}$$

$$= \frac{\mathbb{E}(p(\pi^{\text{Db++}}_{[n+2]} \mid \pi^{\text{Db+}}_{[n+1]}, X_1, ..., X_{n+1}, h_p, \mathbf{p}) \mid \Pi^{\text{Db+}}_{[n+1]} = \pi^{\text{Db+}}_{[n+1]}, \mathbf{p})}{\mathbb{E}(p(\pi^{\text{Db++}}_{[n+2]} \mid \pi^{\text{Db+}}_{[n+1]}, X_1, ..., X_{n+1}, h_d, \mathbf{p}) \mid \Pi^{\text{Db+}}_{[n+1]} = \pi^{\text{Db+}}_{[n+1]}, \mathbf{p})} \tag{7.15}$$

$$= \frac{1}{\mathbb{E}(p_{X_{n+1}} | \Pi^{\text{Db+}}_{[n+1]} = \pi^{\text{Db+}}_{[n+1]}, \mathbf{p})}. \tag{7.16}$$

Notice that, in the rare type case, X_{n+1} is observed only once among the X_1, ..., X_{n+1}. Hence, we call it a singleton. Let s_1 denote the number of singletons, and $\mathcal{S}$ the set of indexes of singletons observations in the database. Notice also that the knowledge of $\mathbf{p}$ and $\pi^{\text{Db+}}_{[n+1]}$, is not enough to observe $X_1, ..., , X_{N+1}$. On the other hand, given $\pi^{\text{Db+}}_{[n+1]}$, both s_1 and $\mathcal{S}$ are fixed and known. Given $\mathbf{p}$ and $\pi^{\text{Db+}}_{[n+1]}$, it holds that the distribution of X_{n+1} is the same as the distribution of all other singletons. This implies that:

$$s_1 \mathbb{E}(p_{X_{n+1}} | \pi^{\text{Db+}}_{[n+1]}, \mathbf{p}) = \mathbb{E}(\sum_{i \in \mathcal{S}} p_{X_i} | \pi^{\text{Db+}}_{[n+1]}, \mathbf{p}).$$

Let us denote as X^*_1, .., X^*_K the K different values taken by X_1, ..., X_{n+1}, ordered according to the frequency of their values. Stated otherwise, if n_i is the frequency of x^*_i among $x_1, ..., x_{n+1}$, then $n_1 \geq n_2 \geq ... \geq n_K$. Moreover, in case X^*_i and X^*_j have the same frequency ($n_i = n_j$), then they are ordered according to their values. For instance, if $X_1 = 2$, $X_2 = 4$, $X_3 = 2$, $X_4 = 4$, $X_5 = 3$, $X_6 = 3$, $X_7 = 10$, $X_8 = 13$, $X_9 = 5$, $X_{10} = 4$, $X_{11} = 9$, then $X^*_1 = 4, X^*_2 = 2, X^*_3 = 3, X^*_4 = 5, X^*_5 = 10, X^*_6 = 13$.

By definition, it holds that

$$\mathbb{E}(\sum_{i \in \mathcal{S}} p_{X_i} | \pi^{\text{Db+}}_{[n+1]}, \mathbf{p}) = \mathbb{E}(\sum_{j:\, n_j = 1} p_{X^*_j} | \pi^{\text{Db+}}_{[n+1]}, \mathbf{p}).$$

Notice that $(n_1, n_2, ..., n_K)$ is a partition of $n+1$, which will be denoted as $\pi^{\text{Db+}}_{n+1}$. In the example, $\pi^{\text{Db+}}_{n+1} = (3, 2, 2, 1, 1, 1, 1)$. Since the distribution of $\sum_{j:\, n_j = 1} p_{x^*_j}$ only depends on $\pi^{\text{Db+}}_{n+1}$,

the latter can replace $\pi^{\text{Db+}}_{[n+1]}$. Thus, it holds that

$$\text{LR}_{|\mathbf{p}} = \frac{s_1}{\mathbb{E}(\sum_{j:\, n_j=1} p_{X_j^*} | \pi^{\text{Db+}}_{n+1}, \mathbf{p})}. \tag{7.17}$$

For the same reason explained above, the knowledge of $\mathbf{p}$ and $\pi^{\text{Db+}}_{n+1}$ is not enough to observe $X_1^*, ..., X_K^*$. A more compact representation for $\pi^{\text{Db+}}_{n+1}$ can be obtained by using two vectors $\mathbf{a}$ and $\mathbf{r}$ where a_j are the distinct numbers occurring in the partition, ordered, and each r_j is the number of repetitions of a_j. J is the length of these two vectors, and it holds that $n+1 = \sum_{j=1}^{J} a_j r_j$. In the example above we have that $\pi^{\text{Db+}}_{n+1}$ can be represented by $(\mathbf{a}, \mathbf{r})$ with $\mathbf{a} = (1, 2, 3)$ and $\mathbf{r} = (4, 2, 1)$.

There is a function, χ, treated here as latent variable, which assigns all DNA types, ordered according to their frequency in Nature, to one of the number $\{1, 2, ..., J\}$ corresponding to the position in $\mathbf{a}$ of its frequency in the sample, or to 0 if the type if not observed. Stated otherwise,

$$\chi : \{1, 2, ...\} \longrightarrow \{1, 2, ..., J\}.$$

$$\chi(i) = \begin{cases} 0 & \text{if the } i\text{th most common species is not observed in the sample,} \\ j & \text{if the } i\text{th most common species is one of the } r_j \text{ observed } a_j \text{ times in the sample.} \end{cases}$$

Given $\pi^{\text{Db+}}_{n+1} = (\mathbf{a}, \mathbf{r})$, χ must satisfy the following conditions:

$$\sum_{i=1}^{\infty} \mathbf{1}_{\chi(i)=j} = r_j, \qquad \forall j. \tag{7.18}$$

The map χ can be represented by a vector $\boldsymbol{\chi} = (\chi_1, \chi_2, ...)$ such that $\chi_i = \chi(i)$. In the example above we have that $\boldsymbol{\chi} = (0, 2, 2, 3, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, ..., 0)$.

Notice that, given $\pi^{\text{Db+}}_{n+1} = (\mathbf{a}, \mathbf{r})$, the knowledge of $\boldsymbol{\chi}$ implies the knowledge of X_1^*, ..., X_K^*: indeed it is enough to sort the positive values among the χ_i and take their positions in $\boldsymbol{\chi}$, and solving ties by considering the positions themselves (if $\chi_i = \chi_j$, than the order is given by i and j). For instance, in the example, if we sort the values of $\boldsymbol{\chi}$ and we collect their positions we get $(4, 2, 3, 5, 10, 13)$: the reader can notice that we got back to $X_1^*, ..., X_6^*$.

This means that to obtain the distribution of $X_1^*, ..., X_K^* | \pi^{\text{Db+}}_{n+1}, \mathbf{p}$, which appears in (7.17), it is enough to obtain the distribution of $\boldsymbol{\chi} | \pi^{\text{Db+}}_{n+1}, \mathbf{p}$, and since we are only interested in the mean of the sum of singletons in samples of size $n+1$ from the distribution of $X_1^*, ..., X_K^* | \pi^{\text{Db+}}_{n+1}, \mathbf{p}$, we can just simulate samples from the distribution of $\boldsymbol{\chi} | \pi_{n+1}, \mathbf{p}$ and sum the p_a such that $\chi_a = 1$.

To simulate samples from the distribution of $\boldsymbol{\chi} | \pi_{n+1}, \mathbf{p}$ we use a Metropolis-Hastings algorithm, on the space of the vectors $\boldsymbol{\chi}$ satisfying condition (7.18). Notice that for the model we assumed $\mathbf{p}$ to be infinitely long, but for simulations we will use a finite $\bar{\mathbf{p}}$, of length m. This is equivalent to assume that only m elements in the infinite $\mathbf{p}$ are positive, and the remaining infinite tail is made of zeros. Then the state space of the Metropolis-Hastings Markov

chain is made of all vectors of length m whose elements belong to $\{0, 1, ..., J\}$, and satisfy the condition (7.18). If we start with a initial point $\boldsymbol{\chi}_0$ which satisfies (7.18) and, at each allowed move of the Metropolis-Hastings, we swap two different values χ_a and χ_b inside the vector, condition (7.18) remains satisfied. The algorithm is based on a similar one proposed in Anevski et al. (2013).

This method allows us to obtain the 'true' LR when the vector $\mathbf{p}$ is known. This is rarely the case, but we can put ourselves in a fictitious world where we know $\mathbf{p}$, and compare the true values for the LR with the one obtained by applying our model when $\mathbf{p}$ is unknown. This will be done in the forthcoming section.

7.6 Analysis on a real database

In this section we present the study we made on a database of 18,925 Y-STR 23-loci profiles from 129 different locations in 51 countries in Europe (Purps et al., 2014)[1]. Different analyses are performed by considering only 7 Y-STR loci (DYS19, DYS389 I, DYS389 II, DYS3904, DYS3915, DY3926,DY3937) but similar results have been observed with the use of 10 loci.

First, we calculated the maximum likelihood estimators α_{MLE} and θ_{MLE} using the entire database. Their values are $\alpha_{\text{MLE}} = 0.5$ and $\theta_{\text{MLE}} = 216$.

In order to check if the two-parameter Poisson Dirichlet prior is a sensible choice we first compare the ranked frequencies from the database with the relative frequencies of several samples of size n obtained from realisations of $\text{PD}(\alpha_{MLE}, \theta_{MLE})$. The asymptotic behaviour described in (7.11) is also discussed. Lastly, we will analyse the loglikelihood function for the hyperparameters, given the data $\pi^{\text{Db}+}_{[n+1]}$, in order to a perform a data driven choice for the hyperprior.

7.6.1 Model fitting

In Figure 7.5, the ranked frequencies from the database are compared to the relative frequencies of samples of size n obtained from several realizations of $\text{PD}(\alpha_{MLE}, \theta_{MLE})$. To do so we run several times the Chinese Restaurant seating plan (up to $n = 18,925$ customers): each run is equivalent to generate a new realization $\mathbf{p}$ from the $\text{PD}(\alpha_{MLE}, \theta_{MLE})$. The partition of the customers into tables is the same as the partition obtained from an i.i.d. sample of size n from $\mathbf{p}$. The ranked relative sizes of each table (thin lines) are compared to the ranked frequencies of our database (thick line). One can see that for the most common haplotypes (left part of the plot) there is some discrepancy. However, we are interested in rare haplotypes, which typically have a frequency belonging to the right part of the plot. In that region the two-parameter Poisson Dirichlet follows the distribution of the data quite well.

[1]The database has previously been cleaned by Mikkel Meyer Andersen (`http://people.math.aau.dk/~mikl/?p=y23`).

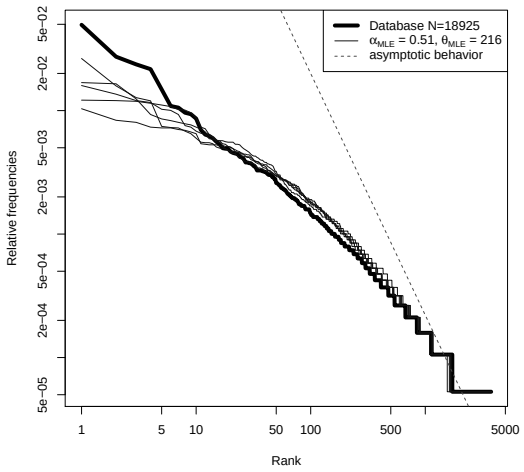


Figure 7.5: Log scale ranked frequencies from the database (thick line) are compared to the relative frequencies of samples of size n obtained from several realizations of $\text{PD}(\alpha_{MLE}, \theta_{MLE})$ (thin lines). Asymptotic power-law behavior is also displayed (dotted line).

The asymptotic behavior described in (7.11) is shown in Figure 7.5 with the dotted line. In the limit over i the thin curves are expected to bend to follow that line. This is not what we observe, but we saw from further simulation studies that we should not expect this power-law behaviour to hold at this sample size for such a big value of θ.

7.6.2 Loglikelihood

It is also interesting to investigate the shape of the loglikelihood function for α and θ given $\pi^{\text{Db++}}_{[n+1]}$. It is defined as

$$l_{n+1}(\alpha, \theta) := \log p(\pi^{\text{Db++}}_{[n+1]} | \alpha, \theta).$$

In Figure 7.6 the loglikelihood reparametrized using $\phi = n\dfrac{1-\alpha}{n+1+\theta}$, and θ instead of α and θ, is displayed. The Gaussian distribution is also displayed (in dashed lines). This is not done to show an asymptotic property, but to show the simmetry of the loglikelihood, which allows to approximate $\mathbb{E}(\Phi \mid \Pi^{\text{Db+}}_{[n+1]} = \pi^{\text{Db+}}_{[n+1]})$ with the marginal mode Φ_{MLE}, if the prior $p(\phi, \theta)$ is flat around $(\phi_{MLE}, \theta_{MLE})$, since it holds that $p(\phi, \theta \mid \pi^{\text{Db+}}_{[n+1]}) \propto l_{n+1}(\phi, \theta) \times p(\phi, \theta)$.

Hence, one can approximate the LR itself in the following way:

$$\text{LR} \approx \frac{n+1+\theta_{MLE}}{1-\alpha_{MLE}}. \tag{7.19}$$

Notice that this is equivalent to an hybrid approach, in which the parameters are estimated through the MLE (frequentist) and their values are plugged into the Bayesian LR.

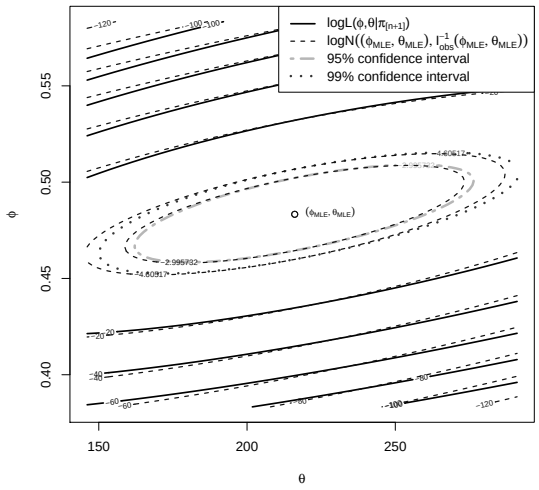


Figure 7.6: Relative loglikelihood for $\phi = n\frac{1-\alpha}{n+1+\theta}$ and θ compared to a Gaussian distribution displayed with 95% and 99% confidence intervals

The Gaussian behavior of Figure 7.6 was unexpected. We expect that increasing n, α and θ would become independent, thus the ellipses will rotate.

7.6.3 Analyzing the error

A real Bayesian statistician chooses the prior and hyperprior according to his beliefs. Depending on the choice of the hyperprior over α and θ he may or may not believe in the approximation (7.19), but he does not really talk of 'error'. However, hardliner Bayesian statisticians are a rare species, and most of the time the Bayesian procedure consists in choosing priors (and hyperpriors) which are a compromise between personal beliefs and mathematical convenience. It is thus interesting to investigate how good it is the choice of such priors. This can be done by comparing the Bayesian likelihood ratio with the likelihood ratio a frequentist would obtain if the vector $\mathbf{p}$ was known, and for the same reduction of data. This is what we call 'error': in other words, at the moment we are considering the Bayesian nonparametric method proposed in this paper as a way to estimate (notice the frequentist terminology) the true $\mathrm{LR}_{|\mathbf{p}}$. If we denote by p_x the population proportion of the matching profile, another interesting comparison is the one between the Bayesian likelihood ratio and the frequentist likelihood ratio $1/p_x$ (here denoted as LR_f) that one would obtain knowing $\mathbf{p}$, but not reducing the data to partition. This is a sort of benchmark comparison, and tells us how much we lose by using the Bayesian nonparametric methodology, and by reducing data. In order to evaluate how much we lose due to the sole reduction of the data, one can compare $\mathrm{LR}_{|\mathbf{p}}$ with LR_f. In total there are three quantities of interest ($\log_{10}\mathrm{LR}$, $\log_{10}\mathrm{LR}_{|\mathbf{p}}$, and $\log_{10}\mathrm{LR}_f$), and three differences of interest, which will be denoted as

- $\mathrm{Diff}_1 = \log_{10}\mathrm{LR} - \log_{10}\mathrm{LR}_{|\mathbf{p}}$
- $\mathrm{Diff}_2 = \log_{10}\mathrm{LR} - \log_{10}\mathrm{LR}_f$

- $\text{Diff}_3 = \log_{10} \text{LR}_f - \log_{10} \text{LR}_{|\mathbf{p}}$

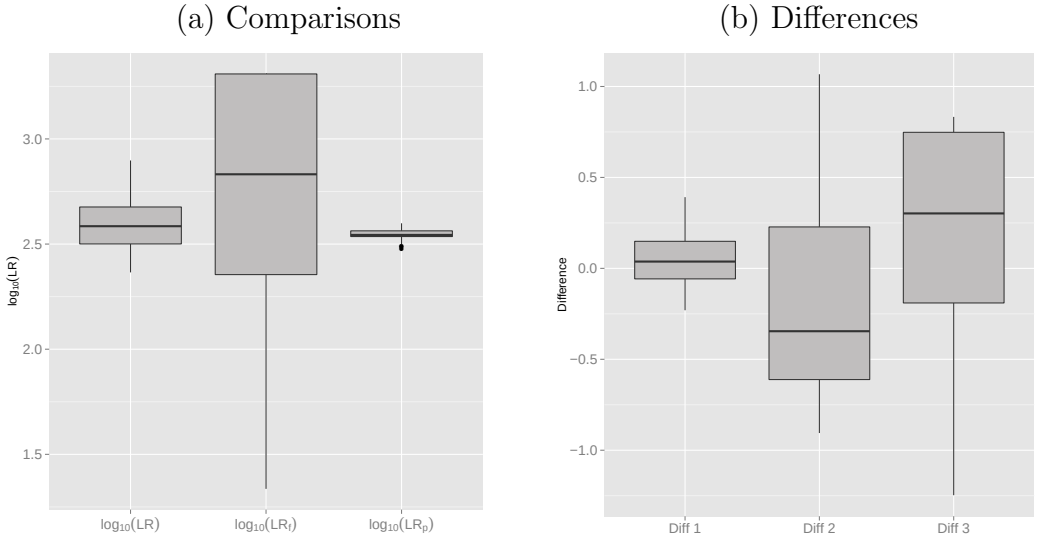


Figure 7.7: (a) comparison between the distribution of $\log_{10} \text{LR}$ and $\text{LR}_{|\mathbf{p}}$. (b) the error $\log_{10} \text{LR}_{|\mathbf{p}} - \log_{10} \text{LR}$.

In order to make the computational effort feasible, instead of using the big database of Purps et al. (2014), we consider the hapolotype frequencies for the sole Dutch population (of size 2037), and we pretend that they are the frequencies from the entire population of possible perpetrators, and we simulate the distribution of the three likelihood ratios of interest.

In Table 7.1 and Figure 7.7 (left part) we compare the distribution of $\log_{10}(\text{LR}_{|\mathbf{p}})$, $\log_{10} \text{LR}$, and $\text{LR}_{|\mathbf{f}}$ obtained by 100 samples of size 100 from this population. The Metropolis-Hastings algorithm explained in Section 7.5.1 can be used to obtain $\text{LR}_{|\mathbf{p}}$.

The distribution of the benchmark likelihood ratio ($\log_{10}(\text{LR}_f)$ has more variation than the distribution of the Bayesian likelihood ratio, while $\log_{10}(\text{LR}_{|\mathbf{p}})$ appears to be the most concentrated around its mean. This is probably due to the small size of the population and of the sampled databases.

In Table 7.2 and Figure 7.7 (right part) we consider the distribution of the three differences, as defined above. Diff_1 is the smallest and the most concentrated: it ranges between -0.4 and 0.23 and has a small standard deviation. It means that the nonparametric Bayesian likelihood ratio obtained as in (7.19) can be thought of as a good approximation of the frequentist likelihood ratio for the same reduction of data ($\log_{10} \text{LR}_{|\mathbf{p}}$). This difference has three components: the approximation (7.19), the MLE estimation of the hyperparameters, and the choice of a prior distribution (two-parameter Poisson Dirichlet) which is quite realistic, as shown in Figure 7.5, but not perfectly fitting the actual population. Moreover, $\log_{10}(\text{LR}_{|\mathbf{p}})$ is not derived exactly, but is obtained using the Metropolis Hasting approximation.

Notice that the difference increases if the Bayesian nonparametric likelihood is compared to the benchmark likelihood ratio (Diff_2). However, it ranges between -1 and +1, but most of the time the difference is between about -0.6 and 0.2, thus small.

The analysis of the distribution Diff_3 tells us that reducing data to the partitions implies a

	Min	1st Qu.	Median	Mean	3rd Qu.	Max	sd
$\log_{10} \mathrm{LR}$	2.365	2.501	2.585	2.596	2.676	2.897	0.116
$\log_{10} \mathrm{LR}_{\lvert \mathbf{p}}$	2.476	2.536	2.543	2.547	2.563	2.599	0.024
$\log_{10} \mathrm{LR}_f$	1.336	2.355	2.832	2.794	3.309	3.309	0.481

Table 7.1: Summaries of the distribution of $\log_{10} \mathrm{LR}$, $\log_{10}(\mathrm{LR}_{\lvert \mathbf{p}})$, and $\log_{10} \mathrm{LR}_f$.

	Min	1st Qu.	Median	Mean	3rd Qu.	Max	sd
Diff_1	-0.23	-0.058	0.037	0.049	0.149	0.391	0.134
Diff_2	-0.905	-0.611	-0.346	-0.198	0.228	1.067	0.492
Diff_3	-1.247	-0.19	0.302	0.247	0.748	0.833	0.484

Table 7.2: Summaries of the distribution of Diff_1, Diff_2, and Diff_3.

loss in the capability to discriminate between the competing hypotheses of at most one order of magnitude, thus not terribly bad.

7.7 Conclusion

This paper discusses the first application of a Bayesian nonparametric method to likelihood ratio assessment in forensic science, in particular to the challenging situation of the rare type match. If compared to traditional Bayesian methods such as those described in Cereda (2016a), it presents many advantages. First of all, the prior chosen for the parameter $\mathbf{p}$ is more realistic for the population whose frequencies we want to model. Moreover, although the theoretical background on which it lies may seem very technical and difficult, the method is extremely simple to apply for practical use, thanks to the discussed approximation: indeed, simulation experiments show that an hybrid empirical approach is justified, at least using Y-STR data from European populations. The likelihood ratio obtained with this method is also compared to the frequentist likelihood ratio obtained, knowing the population frequencies of each type, both reducing and not reducing the data. The differences are quite small, reaching at most 1 order of magnitude. More could be done in the future: for instance, investigate other nonparametric priors, and use more realistic populations.

Acknowledgment

I am indebted to Jim Pitman and Alexander Gnedin for their help in understanding their important theoretical results, and to Mikkel Meyer Andersen for providing a cleaned version of the database of Purps et al. (2014). This research was supported by the Swiss National Science Foundation, through grants no. 105311-144557 and 10531A-156146, and carried out in the context of a joint research project, supervised by Franco Taroni (University of Lausanne, Ecole des sciences criminelles), and Richard Gill (Mathematical Institute, Leiden University).

Chapter 8

A solution for the rare type match problem when using the DIP-STR marker system

This chapter is based on: Cereda, G., Gill, R. D., and Taroni, F. "A solution for the rare type match problem when using the DIP-STR marker system". Submitted to *Forensic Science International: Genetics*.

Abstract

The rare type match problem is an evaluative challenging situation in which the analysis of a DNA profile reveals the presence of (at least) one allele which is not contained in the reference database. This situation is challenging because an estimate for the frequency of occurrence of the profile in a given population needs sophisticated evaluative procedures.

The rare type match problem is very common when the DIP-STR marker system, which has proven itself very useful for dealing with unbalanced DNA mixtures, is used, essentially due to the limited size of the available database. The object-oriented Bayesian network proposed in Cereda et al. (2014b) to assess the value of the evidence for general scenarios, was not designed to deal with this peculiar situation. In this paper, the model is extended and partially modified to be able to calculate the full Bayesian likelihood ratio in presence of any (observed and not yet observed) allele of a given profile. The method is based on the approach developed in Cereda (2016a) for Y-STR data. Alternative solutions, such as the plug-in approximation and an empirical Bayesian methodology are also proposed and compared with the results obtained with the full Bayesian approach.

8.1 Introduction

The most common task of a forensic scientist or statistician is to quantify the probative value of the observation of some scientific findings (e.g. DNA profiles, fragment of paint, fibres), under the hypotheses of interest for the court of justice. This is done through the quantification of the likelihood ratio. In case the hypotheses of interest deal with whether the recovered material has the same origin as some control material, it is important to be able to quantify the rarity of the corresponding characteristics. For instance, the evidence can be the correspondence between the DNA profile of a crime stain and of a suspect: the rarer the profile, the more probative is the scientific finding regarding propositions about the source. The rarity of the profile of interest is often used to assign the probability of the random occurrence of the given stain, and some available (and relevant) database is used to support the scientist's assignment.

The 'rare type match problem', also called 'the fundamental problem of forensic mathematics' (Brenner, 2010) is the situation in which the corresponding characteristic has not been observed in the relevant reference database for the case. One example is the DIP-STR marker system, a rather novel genotyping technique, proposed in Castella et al. (2013), which turned out to be very useful to analyse DNA mixtures if the proportion of the DNA quantities of the, say, two contributors is more extreme than 1:10. Due to limited size of available databases, rare DIP-STR profiles are often encountered.

A Bayesian framework for evaluating DIP-STR results was developed in Cereda et al. (2014b), using object-oriented Bayesian networks, with the aim of calculating the likelihood ratio for mixtures of two contributors, when the major contributor's genotype is known and the two competing hypotheses are 'the minor contributor is the suspect' (h_p) and 'the minor contributor is an unknown person, unrelated to the suspect' (h_d), also extended to cases where the suspect is missing.

This paper proposes a Bayesian solution for assigning the likelihood ratio for mixture results in presence of a rare type match, that is when at least one of the DIP-STR alleles of the contributors is not present in the reference database. This situation was not covered by Cereda et al. (2014b).

The Bayesian model adopted is based on a similar one proposed in Cereda (2016a). Several issues concerning Bayesian methodology, and notation, have been improved.

The paper is structured as follows. Section 8.2 discusses the use of the DIP-STR marker system for extremely unbalanced mixtures, while Section 8.3 describes the object-oriented Bayesian network that was built to evaluate DIP-STR profiling results in Cereda et al. (2014b). The chosen notation and the definition of what a full Bayesian approach to likelihood ratio assessment is, can be found in Sections 8.4 and 8.5, respectively. The model developed to evaluate results from mixtures of two contributors in presence of the rare type match problem (described in Section 8.6) is detailed in Sections 8.7 and 8.8. More detailed descriptions of the development of the full Bayesian likelihood ratio, which takes advantage of the Lemma introduced in Section 8.9, are confined to the Appendix. A discussion about the choice of the prior distribution for the parameters is also provided in Section 8.10, while conclusions can be found in Section 8.11.

8.2 DIP-STR marker system for extremely unbalanced mixtures

A DIP-STR marker is a compound marker made of a DIP (Deletion/Insertion polymorphism, Weber et al. (e.g., 2002)), and of a standard STR polymorphism. These two polymorphisms are chosen less than 500 bp apart, in order to be dependent on one another.

Standard methods for the analysis of DNA mixtures, such as STR markers (Butler, 2011), fail to detect the DNA of contributors whose DNA constitutes less than the 10% of the total DNA material (Clayton and Buckleton, 2005). On the other hand, as long as the minor contributor has at a specific locus at least one DIP allele different from the DIP alleles of the suspect, the DIP-STR marker system allows the selected amplification of its DIP-STR genotype, up to mixture proportions as extreme as 1:1000,

At each DIP-STR locus, the possible configurations are the following (summarized in Table 8.1).

- In the case where the major and minor contributors are DIP homozygous with different alleles (i.e., one is L-L and the other is S-S) both DIP-STR alleles of the minor can be detected. This is the best scenario the scientist can be faced with.
- If the major is DIP homozygous (for instance L-L) and the minor is DIP heterozygous, only one of the two DIP-STR alleles of the minor can be detected: the one with the other DIP allele (in the example the allele S).
- The worst situation is the one in which the major is DIP heterozygous, or both contributors are homozygous for the same DIP alleles: in these cases the DNA profile of the minor cannot be obtained.

DIP genotype of major/minor contributor	**DIP-STR alleles observed in the trace**	**Information gained for the second contributor**
Hom/Hom (different kind)	2 (if STR het)	Yes completely
	1 (if STR hom)	Yes
Hom/Het	1 (regardless STR)	Yes
Hom/Hom (same kind)	0 (regardless STR)	No
Het/Hom	0 (regardless STR)	No
Het/Het	0 (regardless STR)	No

Table 8.1: Informativeness of genotypic configurations. 'Hom' denotes homozygous for the DIP allele, and 'Het' heterozygous.

A first panel of 10 DIP-STR markers was presented in Castella et al. (2013). A second panel with 9 additional DIP-STR markers has recently been provided in Oldoni et al. (2015). When one analyses a mixed stain, each of the 19 available markers may present one of the three situations described above.

8.3 Bayesian network for evaluating DIP-STR profiling results from unbalanced DNA mixtures.

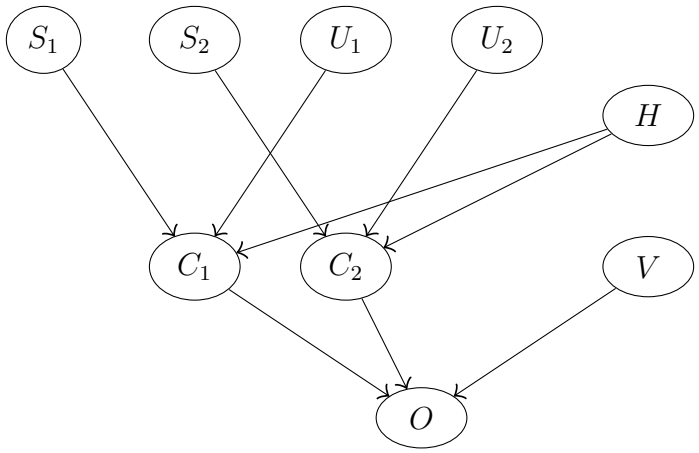


Figure 8.1: Bayesian network corresponding to the object-oriented Bayesian network of Cereda et al. (2014b). The meaning of the nodes is described in Section 8.3.

In Cereda et al. (2014b) a locus specific object-oriented Bayesian network (OOBN), designed to assist the evaluation of DIP-STR results obtained from mixtures with two contributors, is proposed. The network, reproducing the mechanism described in Section 8.2 and in Table 8.1, is proposed here in the form of a Bayesian network (see Figure 8.1). It is suitable for a situation in which the DIP-STR profile of a suspect (potential contributor to the mixture) is available. The two hypotheses of interest are 'the minor contributor is the suspect' (h_p) and 'the minor contributor is an unknown person, unrelated to the suspect' (h_d). The major contributor is often referred to as "the victim", taken as a known contributor, and his/her DIP-STR profile is generally available.

It is important to notice that the only information needed from the known major contributor regards his DIP alleles. Hence, the only node in the network which concerns the victim, V, has three possible states: $HomoL$, $HomoS$ and $Hetero$. The node H represents the two hypotheses of interest defined above.

With the exception of node O, the remaining part of the network deals with the unknown minor contributor. Nodes S_1 and S_2 represent the two DIP-STR alleles of the suspect. Nodes U_1 and U_2 represent the two DIP-STR alleles of the alternative (unknown) contributor in a two-person mixture. Nodes C_1 and C_2 represent the DIP-STR alleles of the actual second contributor (for example, the suspect). Depending on the state of node H, the second contributor's allele can be a copy of S_1 and S_2 (under state h_p), or of U_1 and U_2 (under state h_d). The state of node O, which contains results obtained from the mixture, depends on the combination of V, C_1 and C_2 (according to Table 8.1).

The probability tables for the nodes are of different types. In the scenario considered, node V is observed, since the major contributor is known. As such, its probability table is not relevant for the final result because its state is fixed, thus it is filled with equal prior probabilities

for its three states. The same holds for node H, which is in turn instantiated to obtain the numerator and the denominator of the likelihood ratio.

Nodes C_1 and C_2 are deterministic given H, S_1, S_2, U_1, and U_2: if H is in state h_p, then C_1 and C_2 are copies of, respectively, S_1 and S_2, otherwise they are copies of U_1 and U_2. Also node O is deterministic, given nodes V, H, C_1 and C_2: its probability table is filled out with 0's and 1's (according to the conditions defined in Table 8.1). The states of nodes S_1, S_2, U_1, U_2, C_1, and C_2 are La, Lb, Lx, Sa, Sb, Sx. Notice that at each DIP-STR locus there may be more than six possible alleles: La, Lb, Sa, Sb are used to represent the two alleles that at most could be observed, while Sx and Lx represent all the other (not observed) alleles different from a and b. In Cereda et al. (2014b), this solution was preferred to having the entire list of DIP-STR alleles, in order to make the model simpler, and usable for different loci. The disadvantage is that, at each new case, the meaning of these symbols changes, and the probability tables have to be adapted accordingly. In this paper, we will develop a methodology to overcome this constraint.

The probability tables for nodes S_1, S_2, U_1, and U_2 should be filled with the allelic proportions corresponding to the alleles represented by names La, Lb, etc., in the population of interest. These allelic proportions are unknown, but we a have a database of DIP-STR alleles, which we can consider as a random sample from the population of interest. In Cereda et al. (2014b), a Dirichlet distribution with all parameters equal to one was used as prior for the DIP-STR allelic proportions, and the probability tables for nodes S_1, S_2, U_1, U_2 were filled out with the posterior means (conditional to the observation of the database). However, as discussed in Section 8.5, this approach suffers from some limitations and it can be improved and made more consistent with the Bayesian theory. Moreover, the number of possible distinct DIP-STR alleles was chosen by looking at those in the database. Thus, the model was not suitable to be used when new alleles (not previously detected) were observed. This paper aims at solving these problems.

8.4 Notation

Throughout the paper the following notation is chosen: random variables and their values are denoted, respectively, with uppercase and lowercase characters: x is a realization of X. Random vectors are denoted with bold characters: $\mathbf{x}$ is a realization of the random vector $\mathbf{X}$. Probability is denoted with $\Pr(\cdot)$, while density of a continuous random variable X is denoted alternatively by $p_X(x)$ or by $p(x)$ when the subscript is clear from the context. For a discrete random variable Y, the density notation $p_Y(y)$ and the discrete one $\Pr(Y = y)$ will be alternately used. Moreover, we will use shorthand notation like $p(y \mid x)$ to stand for the probability density of Y with respect to the conditional distribution of Y given $X = x$.

Given $k \geq 2$, and $\alpha = (\alpha_1, ..., \alpha_k)$ such that $\alpha_i > 0$,

$$\mathbf{X} \sim \mathrm{Dir}^k(\alpha_1, ..., \alpha_k)$$

means that vector $\mathbf{x}$ follows a k-dimensional Dirichlet distribution (Press, 2009), whose den-

sity is

$$p(\mathbf{x}) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} x_i^{\alpha_i - 1}.$$

In the appendix, we will denote with $\mathbf{z} = (\mathbf{x}, y)$ the vector $\mathbf{z}$ obtained by adding element y at the end of vector $\mathbf{x}$.

8.5 Full Bayesian approach

In the case of interest, for each analysed locus the forensic scientist or statistician is given the following input data: the victim's and the suspect's DIP-STR profile (denoted as E_v and E_s respectively), along with the DIP-STR alleles obtained from the mixture (E_m). This data has to be evaluated in the light of the hypotheses of interest (h_p and h_d) as defined in Section 8.1. The evaluation of such evidence heavily depends on the allelic proportions of the DIP-STR alleles of the trace and of the suspect, which are unknown. The vector $\boldsymbol{\theta}$, containing the population proportions of all the possible DIP-STR alleles at the considered locus, is the nuisance parameter of the model. A database (denoted here as D), consisting of a list of DIP-STR alleles from the population of interest is given to the statistician, in order to support him in the assessment of the uncertainty about $\boldsymbol{\theta}$. The data to evaluate are thus made of E=(E_v, E_s, E_m) and D. This notation reflects the distinction described in Cereda (2016a) between 'evidence', data directly related to the crime, and 'background', data related only to the nuisance parameter of the model.

The full Bayesian approach consists of modelling all these variables, including $\boldsymbol{\theta}$, as random variables whose joint distribution Pr reflects prior belief of the expert.

The largely accepted method to evaluate the data in order to discriminate between the two hypotheses of interest, is the calculation of the *Bayes factor* (BF), in forensic context regularly called *likelihood ratio* (LR). It is defined as the ratio of the probabilities of observing the data under the two competing hypotheses:

$$\mathrm{LR} = \frac{\Pr(E = e, D = d \mid H = h_p)}{\Pr(E = e, D = d \mid H = h_d)} = \frac{\Pr(E = e \mid D = d, H = h_p)}{\Pr(E = e \mid D = d, H = h_d)}, \tag{8.1}$$

where the last equality holds in virtue of the independence of database and hypotheses.

The nuisance parameter $\boldsymbol{\theta}$ has been integrated out according to its prior distribution. Notice indeed that $\boldsymbol{\theta}$ does not appear in (8.1).

In Cereda et al. (2014b), we used a different approach: Bayesian estimates of the allelic proportions were plugged into the probability tables for nodes S_1, S_2, U_1, and U_2. This is equivalent to using a likelihood ratio for a given $\boldsymbol{\theta}$, such as

$$\mathrm{LR} = \frac{\Pr(E = e \mid \boldsymbol{\Theta} = \boldsymbol{\theta}, D = d, H = h_p)}{\Pr(E = e \mid \boldsymbol{\Theta} = \boldsymbol{\theta}, D = d, H = h_d)},$$

and to plug inside the estimates for $\boldsymbol{\theta}$.

The plug-in method can be seen as an approximation to the full Bayesian method (Cereda, 2016a). To obtain it, a Bayesian network is built which allows one to use an integrated full Bayesian approach, by introducing, among others, a node which represents the database D, and a node that represents the nuisance parameter $\boldsymbol{\theta}$. The full Bayesian approach is then compared to the plug-in method, to check the impact of the approximations.

8.6 Rare type match problem

When the findings to evaluate include a correspondence between the DNA profile of a particular piece of evidence (i.e., a trace of unknown origin) and a suspect's DNA profile, but at least one of the alleles of this profile are not present in the available database, it is difficult to assess the uncertainty over the population proportion of that allele. It is likely to be a rare allele (from which the term *rare type match problem*) but it is challenging to quantify how rare. This assessment is important for the quantification of the likelihood ratio: the rarer the matching profile, the larger is the likelihood ratio.

Using DIP-STR data, it is very likely to encounter the rare type match problem, because the available database size is still limited (Oldoni et al., 2015). The same happens when Y-chromosome (or mitochondrial) DNA profiles are used, since the set of possible Y-STR profiles is extremely large. As a consequence, most of the Y-STR haplotypes are not represented in the database. In Cereda (2016a,b,c) several (Bayesian and frequentist) solutions are proposed for the rare type match problem for Y-STR data. The object-oriented Bayesian network of Cereda et al. (2014b), here presented in Figure 8.1, cannot be used in the case of a rare type match problem: there, the number of different alleles at a given locus was considered as fixed, and equal to that observed in the database. This makes that model useless in cases where new DIP-STR alleles are observed.

As a solution, we will consider the number of different DIP-STR alleles present in the population as random, by introducing additional variables in the model, explained in detail in Section 8.7. This is based on one of the Bayesian methods proposed in Cereda (2016a).

8.7 A prior for $\boldsymbol{\theta}$

Let us denote with L-STR (or S-STR) the DIP-STR alleles which have the DIP part equal to L (or S). Assume that, at a specific locus, there are at most m theoretically possible L-STR alleles and m theoretically possible S-STR alleles. The random vector $\boldsymbol{\Theta} = (\Theta_1^L, ..., \Theta_m^L, \Theta_1^S, ..., \Theta_m^S)$ contains the population proportions of all the potential $2m$ DIP-STR alleles at that locus (for instance, alphabetically ordered).

Only k^L (k^S) of the m possible L-STR (S-STR) alleles are actually present in nature (or more specifically in the population of interest), but k^L and k^S are unknown. Which of the m L-STR alleles are those k^L and k^S is not known either.

The vector $\mathbf{t}^L$ contains the ordered positions (from 1 to m) of the k^L L-STR alleles present

in the population of interest. $\mathbf{t}^L$ is modelled through a random variable $\mathbf{T}^L$: each possible configuration $\mathbf{t}^L$ is assumed as equiprobable, hence it is chosen uniformly at random from the possible $\binom{m}{k^L}$ configurations. Random vector $\mathbf{T}^S$ is defined similarly. Notice that $\theta_i^L = 0, \forall i \notin \mathbf{t}^L$, and $\theta_i^S = 0, \forall i \notin \mathbf{t}^S$.

Specifying $\boldsymbol{\Theta}$ is equivalent to specifying three random variables $\boldsymbol{\Phi}^L$, $\boldsymbol{\Phi}^S$, Ψ. Ψ is the sum of the occurrence probabilities of the L-STR alleles

$$\psi = \sum_{i=1}^{m} \theta_i^L,$$

while $\boldsymbol{\phi}^L$ is the normalized vector of the occurrence probabilities of the L-STR alleles. Stated otherwise,

$$\boldsymbol{\phi}^L = (\frac{\theta_1^L}{\psi}, ..., \frac{\theta_m^L}{\psi}).$$

Similarly, $\boldsymbol{\phi}^S$ is the normalized vector of the frequencies of the S-STR alleles:

$$\boldsymbol{\phi}^S = (\frac{\theta_1^S}{1-\psi}, ..., \frac{\theta_m^S}{1-\psi}).$$

The prior distribution for $\boldsymbol{\theta}$ can be described in terms of the prior over $\boldsymbol{\Phi}^L$, $\boldsymbol{\Phi}^S$, and Ψ, which will be taken to be independent. The latter is distributed according to a Beta(1,1), while the positive entries of $\boldsymbol{\phi}^L$, i.e., $(\phi_i^L \mid i \in \mathbf{t}^L)$ are Dirichlet distributed, given $\mathbf{t}^L$, with all hyperparameters equal to α. The same holds for $\boldsymbol{\phi}^S$ given $\mathbf{t}^S$. Hence, the distribution of node $\boldsymbol{\theta}$ can be described in terms of the distribution of seven additional random variables, whose conditional dependencies can be described by the Bayesian network of Figure 8.2. Bayesian networks using beta and Dirichlet distributions in forensic contexts are presented in Biedermann et al. (2011a). Other examples can also be found in Taroni et al. (2014).

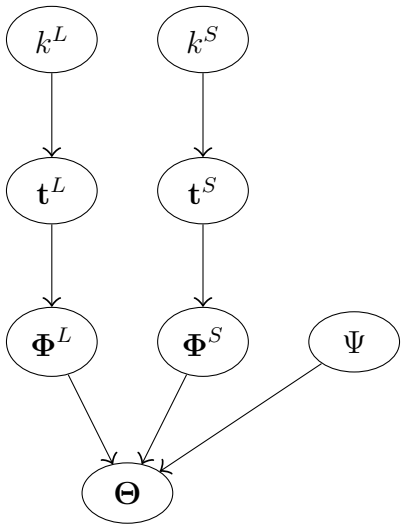


Figure 8.2: The conditional dependency relationships of the random variables used to build the distribution of $\boldsymbol{\theta}$. The definition of the nodes can be found in Section 8.7.

8.8 Full model

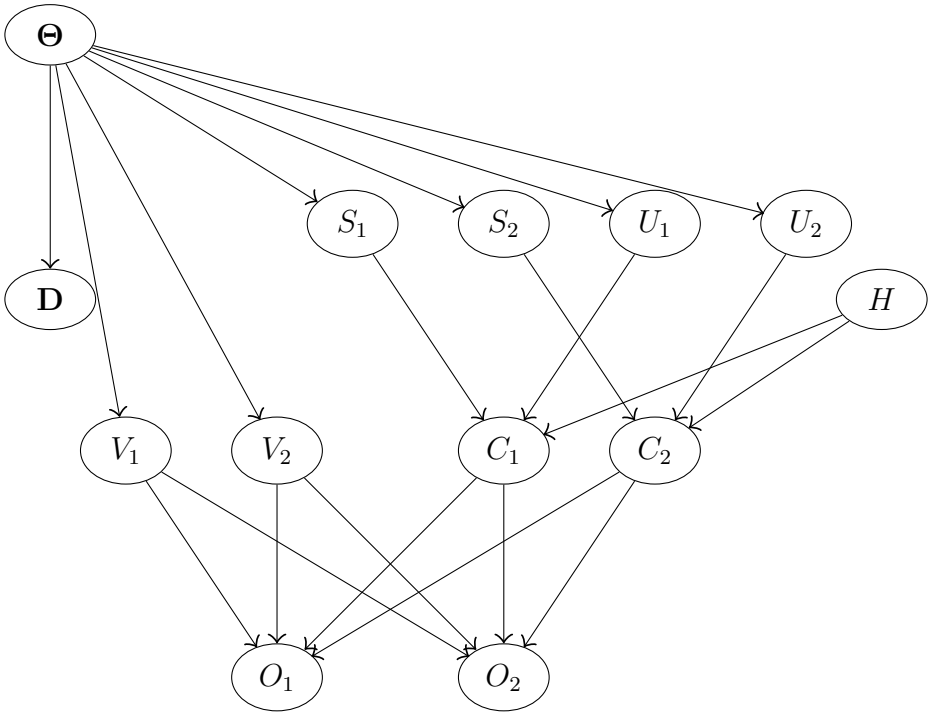


Figure 8.3: Bayesian network for the Dirichlet-multinomial model with a random number of types, to be used for DIP-STR data. The definition of the nodes can be found in Section 8.8.

This model is represented by the Bayesian network of Figure 8.3. Notice that there are differences from the model depicted in Figure 8.1, among which is the presence of node $\boldsymbol{\theta}$ distributed as described in Section 8.7. The first difference lies in the definition of nodes S_1, S_2, U_1, U_2, C_1, and C_2. Their values are couples (L,i) or (S,i) where $i \in \{1, ..., m\}$, describing the position in $\boldsymbol{\theta}$ of the corresponding DIP-STR allele. The same holds for nodes V_1 and V_2, which replace node V of Figure 8.1, and represent the two DIP-STR alleles of the victim. All these nodes are now linked to node $\boldsymbol{\Theta}$ because, given $\boldsymbol{\Theta} = \boldsymbol{\theta}$, the random variables S_1, S_2, U_1, U_2, V_1, V_2 have the following density (with parameter $\boldsymbol{\theta}$):

$$p((j, i) \mid \boldsymbol{\theta}) = \theta_i^j, \quad \forall j \in \{L, S\}, \forall i \in \{1, ..., m\}. \tag{8.2}$$

The second difference is the presence of two nodes O_1 and O_2, instead of a single node O as in Figure 8.1. O_1 represents one of the DIP-STR alleles observed from the mixture (if any, 0 otherwise). O_2 is always 0 unless we are in the situation described by the first row of Table 8.1, where two DIP-STR alleles are observed. In this case, the convention is for O_1 and O_2 to be ordered alphabetically.

The random vector $\mathbf{D}$ represents the available database of size n, through the list of labels ((L, i) or (S, i)) of the DIP-STR alleles contained in the database. The order does not matter, so we can choose the order in which the alleles appear in the database. A particular configuration of $\mathbf{D}$ is denoted as $\mathbf{d} = (d_1, ..., d_n)$, where, given $\boldsymbol{\theta}$, each component is i.i.d. with the same density as in (8.2).

According to this notation, the likelihood ratio for the scenario of interest can be written as

$$\text{LR} = \frac{p(o_1, o_2, s_1, s_2, v_1, v_2, \mathbf{d} \mid h_p)}{p(o_1, o_2, s_1, s_2, v_1, v_2, \mathbf{d} \mid h_d)} = \frac{p(o_1, o_2 \mid s_1, s_2, v_1, v_2, \mathbf{d}, h_p)}{p(o_1, o_2 \mid s_1, s_2, v_1, v_2, \mathbf{d}, h_d)}. \tag{8.3}$$

Due to the complexity of the chosen distribution, the Bayesian network cannot be treated with available software, such as Hugin, or OpenBUGS. However, the likelihood ratio can be obtained analytically using the Lemma presented in Section 8.9.

8.9 Lemma

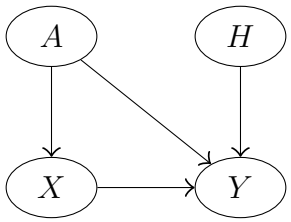


Figure 8.4: Conditional dependencies of the random variables of the Lemma

Lemma 3. *Given four random variables A, H, X and Y, whose conditional dependencies are represented by the Bayesian network of Figure 8.4, the likelihood function for h, given $X = x$ and $Y = y$ satisfies*

$$\text{lik}(h \mid x, y) \ \propto \mathbb{E}(p(y \mid x, A, h) \mid X = x).$$

This Lemma, proven in Cereda (2016c), is very general: it applies to every group of random variables whose conditional dependencies are represented by the Bayesian network of Figure 8.4, and it is very useful due to the possibility of applying it to a very common forensic situation: the prosecution and the defence disagree on the distribution of part of the data (Y) but agree on the distribution of the other part (X), when the distribution of X and Y depends on some parameters (A). This Lemma can also be used for the DIP-STR model presented in Section 8.7. However, it is not straightforward to identify in the Bayesian network of Figure 8.3 the required structure shown in Figure 8.4. Luckily, the same model can be represented in several ways: we will propose a modification of the Bayesian network of Figure 8.3 into something which more clearly shows the required structure. This will be done in two steps: first, we will remove unnecessary nodes, and then we will group some of the others.

Step 1. The Bayesian network presented in Figure 8.5 is obtained by removing from the Bayesian network of Figure 8.3 nodes U_1, U_2, C_1, and C_2. The conditional probability tables of nodes O_1 and O_2 can be directly expressed in terms of S_1, S_2, V_1, V_2, and H, in a way that makes the model equivalent to the previous one (Figure 8.5).

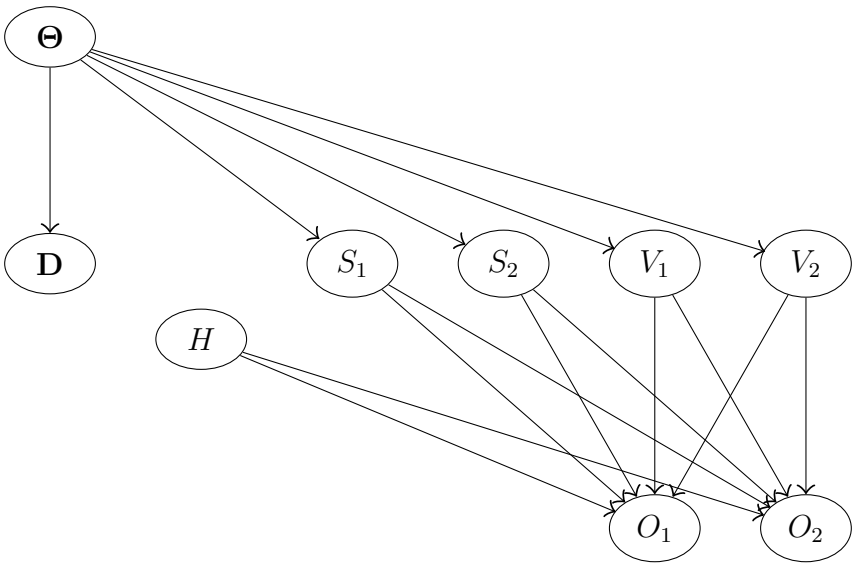


Figure 8.5: An alternative representation of the DIP-STR mixture model presented in Figure 8.3.

Step 2. The Bayesian network of Figure 8.6 can be obtained by substituting some of the nodes of the Bayesian network of Figure 8.5 with a single node. Indeed, instead of having the random vector $\mathbf{D}$ and four additional random variables (S_1, S_2, V_1, and V_2), we can group all these together into a random vector $\mathbf{B}$, of length $n + 4$. The first n elements are the labels contained in $\mathbf{D}$, the fourth to last and third to last are the labels in S_1 and S_2, while the second to last and the last are the labels in V_1, and V_2, respectively.

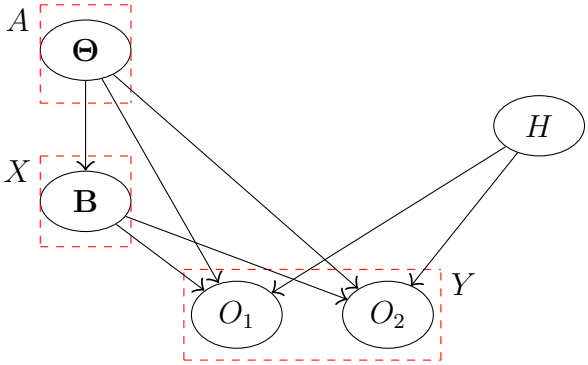


Figure 8.6: A simpler structure for the Bayesian network, suitable to be used for the Lemma. Dashed lines show the choice for the corresponding variables A, X, and Y of Figure 8.4.

The Bayesian network of Figure 8.6 can be used to represent the same model as that represented by Figure 8.1, by carefully adapting the conditional distribution of O_1, and O_2. We can apply the Lemma to our model by defining $Y = (O_1, O_2)$, $X = \mathbf{B}$, and $A = \Theta$. This leads to

$$\mathrm{LR} = \frac{p(o_1, o_2, \mathbf{b} \mid h_p)}{p(o_1, o_2, \mathbf{b} \mid h_d)} = \frac{\mathrm{lik}(h_p \mid o_1, o_2, \mathbf{b})}{\mathrm{lik}(h_d \mid o_1, o_2, \mathbf{b})} = \frac{\mathbb{E}(p(o_1, o_2 \mid \mathbf{b}, \boldsymbol{\Theta}, h_p) \mid \mathbf{B} = \mathbf{b})}{\mathbb{E}(p(o_1, o_2 \mid \mathbf{b}, \boldsymbol{\Theta}, h_d) \mid \mathbf{B} = \mathbf{b})}.$$

Notice that we assume that under the prosecution's hypothesis, $p(o_1, o_2 \mid \mathbf{b}, \boldsymbol{\Theta}, h_p) = 1$.

Therefore, the likelihood ratio can be simplified:

$$\mathrm{LR} = \frac{1}{\mathbb{E}(p(o_1, o_2 \mid \mathbf{b}, \boldsymbol{\Theta}, h_d) \mid \mathbf{B} = \mathbf{b})}. \tag{8.4}$$

Victim's DIP alleles	o_1	o_2	$p(o_1, o_2 \mid \mathbf{b}, \boldsymbol{\Theta}, h_d)$
L-L	(S, i)	(S, j)	$2\Theta_i^S\Theta_j^S$
	(S, i)	0	$(\Theta_i^S)^2 + 2\Theta_i^S\Psi$
	0	0	Ψ^2
S-S	(L, i)	(L, j)	$2\Theta_i^L\Theta_j^L$
	(L, i)	0	$(\Theta_i^L)^2 + 2\Theta_i^L(1-\Psi)$
	0	0	$(1-\Psi)^2$

Table 8.2: Different forms that $p(o_1, o_2 \mid \mathbf{b}, \boldsymbol{\Theta}, h_d)$ can take, based on the DIP-STR alleles observed from the trace and on the victim's DIP alleles. The case in which the victim is heterozygous is not of interest.

$p(o_1, o_2 \mid \mathbf{b}, \boldsymbol{\Theta}, h_d)$ is a function of some components of the vector $\boldsymbol{\Theta}$. The form of this function depends on the combination of the DIP-STR alleles of the victim and of the trace (see Table 8.2). The expectation in the denominator of (8.4) is to be taken using the posterior distribution $\boldsymbol{\Theta} \mid \mathbf{B} = \mathbf{b}$. This is developed in detail in the Appendix.

8.10 Choice of priors

The Appendix shows the form of the denominator of the likelihood ratio for the different cases which one may encounter (for any m, any parameter $\alpha > 0$ for the Dirichlet distribution, and any prior $p(k)$ over k^L and k^S). The choice of a value for α, m, and of a prior over k^L is very delicate. If the expert has strong opinions about the number of L-STR (S-STR) alleles potentially present in nature (m) and in the population of interest (k^L and k^S), he can choose a prior which reflects his beliefs. Otherwise, he can try to use classical priors such as the Poisson distribution, the Negative binomial distribution (both of them truncated so as to have support only over $\{1, ..., m\}$), or the uniform prior over $\{1, ..., m\}$.

8.10.1 Alternative solutions

The most natural choice is to give a uniform prior (over $\{1, ..., m\}$) to k^L and k^S, combined with that of having all the $k^L + k^S$ hyperparameters of the Dirichlet priors over $\boldsymbol{\phi}_{\mathbf{b}}^L$ and $\boldsymbol{\phi}_{\mathbf{b}}^S$ equal one another. These choices represent the lack of knowledge on the number of categories and make the computations tractable.

One of the limitations of having all the hyperparameters α equal one another is that the posterior for k^L, given $\mathbf{b}$ uses only the number of distinct alleles of type L as information, and

ignores other useful information contained in $\mathbf{b}$. An alternative solution, which compensates for this undesired feature, consists of estimating k^L through the database, instead of putting a prior on it. This can be called an empirical Bayesian approach. Notice that such an undesired situation does not appear if personal beliefs are used to specify the prior distribution. Let us define the vector ${\boldsymbol{\phi}_{\mathbf{b}}}^L$ made of the allelic proportions of the L-STR alleles observed in the augmented database, and of a last component $\bar{\phi}^L_{\mathbf{b}}$ which is the sum of the allelic proportions of all the L-STR alleles not observed in $\mathbf{b}$. $\bar{\phi}^L_{\mathbf{b}}$ is the probability of observing a new L-STR allele in the $n+1$th draw from the population. ${\boldsymbol{\phi}_{\mathbf{b}}}^L$ given k^L and $\mathbf{b}$ is Dirichlet distributed, hence we can obtain the posterior expected values of $\bar{\phi}^L_{\mathbf{b}}$:

$$\mathbb{E}(\bar{\phi}^L_{\mathbf{b}} \mid k^L, \mathbf{b}) = \frac{(k^L - k^L_{\mathbf{b}})\alpha}{k^L\alpha + n^L}. \tag{8.5}$$

The so-called *Good-Turing estimator* (Good, 1953) says that the expected value for the probability of the unobserved types can be approximated by the proportion of L-STR singletons (i.e, alleles observed only once) in the database. Stated otherwise,

$$\mathbb{E}(\bar{\phi}^L_{\mathbf{b}} | k^L, \mathbf{b}) \approx \frac{n^L_1}{n^L}. \tag{8.6}$$

where n^L_1 is the number of DIP-STR alleles observed only once in the augmented database. The two quantities (8.5) and (8.6) can be equated in order to obtain an empirical Bayesian estimate of k^L as

$$\hat{k^L} = \frac{n^L_1 n^L + k^L_{\mathbf{b}}\alpha n^L}{\alpha n^L - \alpha n^L_1}.$$

The likelihood ratio for this choice can be obtained using the same formulas developed in the Appendix by using prior over k^L the degenerate prior which gives a probability of one to value $\widehat{k^L}$. This solution allows one to use more information (n^L_1 and n^S_1) from $\mathbf{b}$.

The third option is to use the plug-in approximation proposed by some literature, which estimates the allelic frequencies by their posterior expectation, after the observation of a database. One of the aims of this paper is to investigate the goodness of this approximation, in the case of a rare type match problem.

We did some experiments using marker MID1950-D20S473 (Castella et al., 2013), and considering the two cases described in Table 8.3.

	Victim's alleles	Suspect's alleles	Observed alleles
Case 1	S11 - S11	L2 - L13	L2 - L13
Case 2	S11 - S11	L2 - S12	L2

Table 8.3: Allelic configurations of the victim and of the suspect in the two cases of interest, at marker MID1950-D20S473.

Allele L2 was not contained in the database of reference, hence we are in presence of the rare type match case. The available database, augmented with the victim's and the suspect's DIP-STR alleles, contains 11 different DIP-STR alleles, for a total number of 210 observations (from 105 individuals).

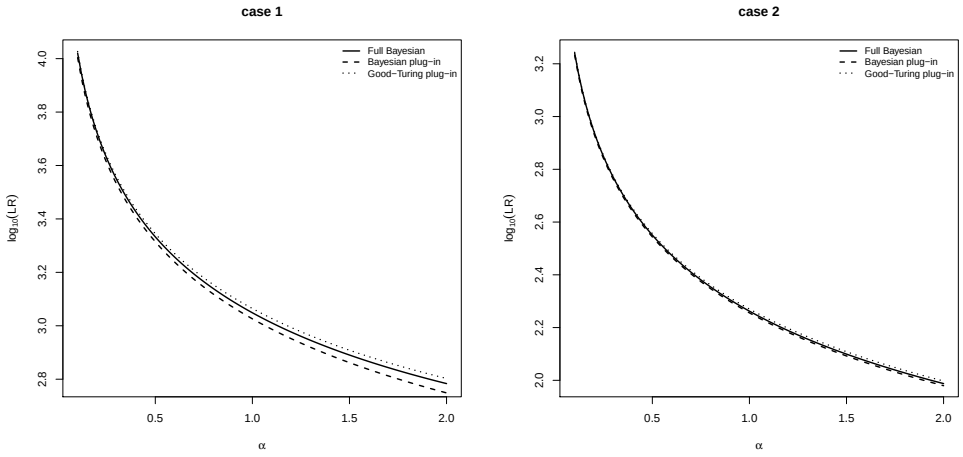


Figure 8.7: Sensitivity analysis for the $\log_{10}(\text{LR})$ obtained with (i) the full Bayesian approach, (ii) the hybrid Good-Turing plug-in (iii) classical Bayesian plug-in, for the two cases described in Table 8.3 when the prior over k^L and k^S is uniform over $\{1, ..., m\}$.

The sensitivity analysis for the $\log_{10}(\text{LR})$, shown in Figure 8.7, has been conducted for different loci and different combinations of alleles, without showing substantial differences (in terms of sensitivity). Moreover, it tells us that the two plug-in approaches represent acceptable solutions in terms of quantification. Varying m does not change Figure 8.7 much.

8.11 Conclusion

Mostly due to the limited size of the available database (about one hundred people in a given relevant population), the rare type match situation is very likely to be encountered when DIP-STR data is used. The recipients of this new technology should be prepared for such an eventuality, which was not taken into account in the OOBN proposed in Cereda et al. (2014b). This paper provides a methodology that allows one to obtain the full Bayesian likelihood ratio also when there are DIP-STR alleles which are not present in the reference database among the alleles of the known contributor and of the suspect. This is done by extending the OOBN, and introducing a more complex prior over the allelic frequencies (a mixture of Dirichlet and uniform distribution) based on a previously developed solution for Y-STR data (Cereda, 2016a). Notice that this issue also represents an opportunity to discuss the use of plug-in approximations which are compared with the full Bayesian likelihood ratio. They proved to be valid approximations.

The sensitivity analysis of the hyperparameters of the prior is also studied. The results show that the likelihood ratio moderately depends on the choices of the parameters α of the Dirichlet prior. Hence, there is the need for further investigations to find better priors, either less sensitive to hyperparameters, or more realistic, such as it was done for Y-STR data in Cereda (2016c).

Appendix. Full Bayesian likelihood ratio development

In Table 8.4, a summary of the relevant symbols used is reported. The aim of this Appendix is to develop the conditional expectation of the functions reported in Table 8.2, which constitute the denominator of the likelihood ratio (8.4). Those conditional expectations can be rewritten in terms of $\boldsymbol{\phi}^L$, $\boldsymbol{\phi}^S$, and ψ, in the following way:

$$\mathbb{E}(2\Theta_i^L\Theta_j^L \mid \mathbf{b}) = \mathbb{E}(2\Phi_i^L\Phi_j^L\Psi^2 \mid \mathbf{b}) = 2\mathbb{E}(\Phi_i^L\Phi_j^L \mid \mathbf{b})\mathbb{E}(\Psi^2 \mid \mathbf{b}),$$
$$\mathbb{E}((\Theta_i^L)^2 + 2\Theta_i^L(1-\Psi) \mid \mathbf{b}) = \mathbb{E}((\Phi_i^L)^2 \mid \mathbf{b})\mathbb{E}(\Psi^2 \mid \mathbf{b}) + 2\mathbb{E}(\Phi_i^L \mid \mathbf{b})\mathbb{E}(\Psi \mid \mathbf{b})\mathbb{E}(1-\Psi \mid \mathbf{b}).$$

The distribution of Ψ given $\mathbf{B}$.

As explained in Section 8.7, $\boldsymbol{\theta}$ can be represented through a set of three independent variables $(\boldsymbol{\phi}^L,\ \boldsymbol{\phi}^S,\ \psi)$. The vector $\mathbf{b}$ can also be reduced by sufficiency to three random variables: $(n^L, \mathbf{n}^L, \mathbf{n}^S)$, where n^L is the total number of observed L-STR alleles in the enlarged database, $\mathbf{n}^L$ is the vector of length m containing the counts in the augmented database of each of the m L-STR alleles, in an order that corresponds to that of $\boldsymbol{\phi}^L$, $\mathbf{n}^S$ is the vector of counts of each of the m S-STR alleles. n^L is binomial distributed with parameters $(n+4,\ \boldsymbol{\psi})$, while $\mathbf{n}^L$ is multinomial distributed with parameters $(n^L, \boldsymbol{\phi}^L)$. Similarly, $\mathbf{n}^S$ is multinomial distributed with parameters $(n^S, \boldsymbol{\phi}^S)$, where $n^S = n + 4 - n^L$ is the number of S-STR alleles in the augmented database.

It holds that the likelihood for $\boldsymbol{\phi}^L, \boldsymbol{\phi}^S$, and ψ factors:

$$p(n^L, \mathbf{n}^L, \mathbf{n}^S \mid \boldsymbol{\phi}^L, \boldsymbol{\phi}^S, \psi) = p(n^L \mid \psi)p(\mathbf{n}^L \mid n^L, \boldsymbol{\phi}^L)p(\mathbf{n}^S \mid n^S, \boldsymbol{\phi}^S).$$

The priors for $\boldsymbol{\phi}^L, \boldsymbol{\phi}^S$, and ψ factors as well, since they are independent. Therefore, the posteriors for $\boldsymbol{\phi}^L, \boldsymbol{\phi}^S$, and for ψ given $\mathbf{b}$ factors as the product of three independent posteriors. Thus, it holds that

$$p(\psi \mid \mathbf{b}) \propto p(n^L \mid \psi)p(\psi),$$

which is a product of the density of a binomial distribution and of a beta prior. By conjugacy,

$$\Psi \mid \mathbf{B} = \mathbf{b} \sim \text{Beta}(1 + n^L, 1 + n^S).$$

In conclusion, by using properties of the Beta distribution, it holds that

$$\mathbb{E}(\Psi^2 \mid \mathbf{b}) = \frac{(n^L+1)(n^L+2)}{(n+6)(n+7)}, \tag{8.7}$$

and

$$\mathbb{E}((1-\Psi)^2 \mid \mathbf{b}) = \frac{(n^S+1)(n^S+2)}{(n+6)(n+7)}. \tag{8.8}$$

Name	Description	Type
m	number of theoretically possible L-STR (and S-STR) alleles	fixed
k^L	number of L-STR allele types present in the population	random
k^S	number of S-STR allele types present in the population	random
$\mathbf{t}^L$	positions (from 1 to m) of the k^L L-STR allele types in the population	random
$\mathbf{t}^S$	positions (from 1 to m) of the k^S S-STR allele types in the population	random
$\boldsymbol{\theta}$	population proportions of the $2m$ possible DIP-STR alleles	random
ψ	sum of the relative frequencies of the L-STR alleles	random
$\boldsymbol{\phi}^L$	normalised vector of the relative frequencies of L-STR alleles	random
$\boldsymbol{\phi}^S$	normalised vector of the relative frequencies of S-STR alleles	random
n	size of the available database	observed
n^L	total number of L-STR alleles in the augmented database	observed
n^S	total number of S-STR alleles in the augmented database	observed
$\mathbf{b}$	labels (j, i) corresponding to each of the $n + 4$ DIP-STR alleles in the augmented database	observed
$\mathbf{b}^L$	labels (L, i) corresponding to each of the n^L L-STR alleles in the augmented database	observed
$\mathbf{b}^S$	labels (S, i) corresponding to each of the n^S S-STR alleles in the augmented database	observed
$k^L_{\mathbf{b}}$	number of distinct L-STR alleles in the augmented database	observed
$k^S_{\mathbf{b}}$	number of distinct S-STR alleles in the augmented database	observed
$\mathbf{n}^L$	counts of all m L-STR alleles in the augmented database	observed
$\mathbf{n}^S$	counts of all m S-STR alleles in the augmented database	observed

Table 8.4: Some relevant symbols used in the paper.

The distribution of ϕ^L and ϕ^S given B.

Let $p(k)$ be the prior distribution over k^L and k^S. In this section we will omit superscripts L and S from k, $\mathbf{t}$, $\boldsymbol{\phi}$, and $\mathbf{n}$, in order to obtain general results valid for both cases. Notice that n will stand for n^L or n^S, and $\mathbf{b}$ will stand for $\mathbf{b}^L$ or $\mathbf{b}^S$ as described in Table 8.4 (so temporarily, the meaning of n, and $\mathbf{b}$ is different from its meaning in the rest of the paper).

Given k, $\mathbf{t}$ is uniformly distributed over the ordered vectors containing k indexes from 1 to m. Let us denote with $k_{\mathbf{b}}$ the number of distinct L-STR (or S-STR) alleles observed in the augmented database, and with $\boldsymbol{\phi}_{\mathbf{b}}$ the vector of length $k_{\mathbf{b}}$ containing only the frequencies of the L-STR alleles observed in the augmented database in the order in which they appear in $\boldsymbol{\phi}$. $\boldsymbol{\phi}_{\mathbf{b}}$ does not sum to one, since there are L-STR alleles of positive frequency, which are not observed: the total probability mass of the unobserved alleles is $\bar{\phi}_{\mathbf{b}} = 1 - \sum_{i=1}^{k_{\mathbf{b}}} \phi_{\mathbf{b}i}$. The vector ${\boldsymbol{\phi}_{\mathbf{b}}}^* = (\boldsymbol{\phi}_{\mathbf{b}}, \bar{\phi}_{\mathbf{b}})$ sums up to one.

We can look for the posterior distribution of ${\boldsymbol{\phi}_{\mathbf{b}}}^*$ given the vector $\mathbf{b}$.

$$p({\boldsymbol{\phi}_{\mathbf{b}}}^* \mid \mathbf{b}) = \sum_k \sum_{\mathbf{t}} p({\boldsymbol{\phi}_{\mathbf{b}}}^* \mid \mathbf{b}, \mathbf{t}) p(\mathbf{t} \mid k, \mathbf{b}) p(k \mid \mathbf{b}) \tag{8.9}$$

It can be proved that

- the posterior density $p({\boldsymbol{\phi}_{\mathbf{b}}}^* \mid \mathbf{b}, \mathbf{t})$ depends on $\mathbf{t}$ only through k. Hence, we can denote it as $p({\boldsymbol{\phi}_{\mathbf{b}}}^* \mid \mathbf{b}, k)$
- if k is less than $k_{\mathbf{b}}$, then $p(k \mid \mathbf{b}) = 0$.
- let us denote with $\mathcal{T}_{k,\mathbf{b}}$ the set of ordered vectors $\mathbf{t}$ of length k and compatible with $\mathbf{b}$ (i.e., which contain among others the positions corresponding to the elements in $\mathbf{b}$). For all the $\mathbf{t}$ which are not in $\mathcal{T}_{k,\mathbf{b}}$, then $p(\mathbf{t} \mid k, \mathbf{b}) = 0$.

We can change the summation indexes in (8.9) to obtain:

$$p({\boldsymbol{\phi}_{\mathbf{b}}}^* \mid \mathbf{b}) = \sum_{k=k_{\mathbf{b}}}^{m} p(k \mid \mathbf{b}) p({\boldsymbol{\phi}_{\mathbf{b}}}^* \mid k, \mathbf{b}) \sum_{\mathbf{t} \in \mathcal{T}_{k,\mathbf{b}}} p(\mathbf{t} \mid k, \mathbf{b}).$$

For any of the $\binom{m-k_{\mathbf{b}}}{k-k_{\mathbf{b}}}$ vectors $\mathbf{t}$ in $\mathcal{T}_{k,\mathbf{b}}$, $p(\mathbf{t} \mid k, \mathbf{b})$ has the same value $\frac{1}{\binom{m-k_{\mathbf{b}}}{k-k_{\mathbf{b}}}}$. Thus, in the end we have that

$$p({\boldsymbol{\phi}_{\mathbf{b}}}^* \mid \mathbf{b}) = \sum_{k=k_{\mathbf{b}}}^{m} p(k \mid \mathbf{b}) p({\boldsymbol{\phi}_{\mathbf{b}}}^* \mid k, \mathbf{b}).$$

The distribution $p(k \mid \mathbf{b})$ can be obtained in the following way.

$$p(k, \mathbf{t}, \boldsymbol{\phi}, \mathbf{b}) = p(k) p(\mathbf{t} \mid k) p(\boldsymbol{\phi} \mid \mathbf{t}) p(\mathbf{b} \mid \boldsymbol{\phi}).$$

Integrating out $\boldsymbol{\phi}$, we obtain

$$p(k, \mathbf{b}, \mathbf{t}) = p(k) p(\mathbf{t} \mid k) \int_{\boldsymbol{\phi}} p(\boldsymbol{\phi} \mid \mathbf{t}) p(\mathbf{b} \mid \boldsymbol{\phi}) \mathrm{d}\boldsymbol{\phi}, \tag{8.10}$$

where the integral contains a Dirichlet density and the categorical density defined in (8.2). They are conjugate, thus we obtain

$$p(k, \mathbf{t} \mid \mathbf{b}) \propto p(k)p(\mathbf{t} \mid k)\frac{\Gamma(k\alpha)}{\Gamma(n+k\alpha)}.$$

Now we can sum over the $\mathbf{t}$ compatible with $\mathbf{b}$, to get to

$$p(k \mid \mathbf{b}) \propto \binom{k}{k_{\mathbf{b}}} p(k)\frac{\Gamma(k\alpha)}{\Gamma(n+k\alpha)}. \tag{8.11}$$

In conclusion,

$$p(\boldsymbol{\phi}_{\mathbf{b}}{}^* \mid \mathbf{b}) \propto \sum_{k=k_{\mathbf{b}}}^{m} \binom{k}{k_{\mathbf{b}}} p(k)\frac{\Gamma(k\alpha)}{\Gamma(n+k\alpha)} p(\boldsymbol{\phi}_{\mathbf{b}}{}^* \mid k, \mathbf{b}), \tag{8.12}$$

where $\boldsymbol{\Phi}_{\mathbf{b}}{}^* \mid K = k, \mathbf{B} = \mathbf{b} \sim \mathrm{Dir}^{k_{\mathbf{b}}+1}(\alpha + \tilde{n}_1, ..., \alpha + \tilde{n}_{k_{\mathbf{b}}}, (k - k_{\mathbf{b}})\alpha)$, and $\tilde{\mathbf{n}}$ is the vector of length $k_{\mathbf{b}}$ with the positive elements of $\mathbf{n}$.

Therefore, (8.12) is a mixture of Dirichlet distributions with weights $w(k) = \binom{k}{k_{\mathbf{b}}} p(k)\frac{\Gamma(k\alpha)}{\Gamma(n+k\alpha)}$. Using properties of the Dirichlet distribution we obtain that, $\forall i, j$ corresponding to different observed DIP-STR alleles:

$$\mathbb{E}(\Phi_i \Phi_j \mid \mathbf{b}) = (\alpha + n_i)(\alpha + n_j)\frac{\sum_{k=k_{\mathbf{b}}}^{m} w(k)g(k)}{\sum_{k=k_{\mathbf{b}}}^{m} w(k)}, \tag{8.13}$$

$$\mathbb{E}(\Phi_i^2 \mid \mathbf{b}) = (\alpha + n_i)(\alpha + n_i + 1)\frac{\sum_{k=k_{\mathbf{b}}}^{m} w(k)g(k)}{\sum_{k=k_{\mathbf{b}}}^{m} w(k)}, \tag{8.14}$$

where $g(k) = \frac{1}{(k\alpha+n)(k\alpha+n+1)}$.

The conditional expectations in Table 8.2

Using (8.7), (8.8), (8.13) and (8.14), we obtain that, $\forall i, j$ corresponding to different observed alleles,

$$\mathbb{E}(\Theta_i^L \Theta_j^L \mid \mathbf{b}) = \mathbb{E}(\Phi_i^L \Phi_j^L \Psi^2 \mid \mathbf{b}) = \mathbb{E}(\Phi_i^L \Phi_j^L \mid \mathbf{b})\mathbb{E}(\Psi^2 \mid \mathbf{b}) \tag{8.15}$$

$$= (\alpha + n_i^L)(\alpha + n_j^L)\frac{\sum_{k=k_{\mathbf{b}}^L}^{m} w^L(k)g^L(k)}{\sum_{k=k_{\mathbf{b}}^L}^{m} w^L(k)}\frac{(n^L+1)(n^L+2)}{(n+6)(n+7)}, \tag{8.16}$$

$$\mathbb{E}((\Theta_i^L)^2 \mid \mathbf{b}) = \mathbb{E}((\Phi_i^L)^2 \Psi^2 \mid \mathbf{b}) = \mathbb{E}((\Phi_i^L)^2 \mid \mathbf{b})\mathbb{E}(\Psi^2 \mid \mathbf{b}) = \tag{8.17}$$

$$= (\alpha + n_i^L)(\alpha + n_i^L + 1)\frac{\sum_{k=k_{\mathbf{b}}^L}^{m} w^L(k)g^L(k)}{\sum_{k=k_{\mathbf{b}}^L}^{m} w^L(k)}\frac{(n^L+1)(n^L+2)}{(n+6)(n+7)}, \tag{8.18}$$

$$\mathbb{E}(\Theta_i^L(1-\Psi) \mid \mathbf{b}) = \mathbb{E}(\Phi_i^L \mid \mathbf{b})(\mathbb{E}(\Psi \mid \mathbf{b}) - \mathbb{E}(\Psi^2 \mid \mathbf{b})) = \tag{8.19}$$

$$= (\alpha + n_i^L) \frac{\sum_{k=k_{\mathbf{b}}^L}^m \frac{w^L(k)}{k\alpha + n^L}}{\sum_{k=k_{\mathbf{b}}^L}^m w^L(k)} \frac{(n^L+1)(n^S+1)}{(n+6)(n+7)}, \tag{8.20}$$

where n and $\mathbf{b}$ have now their original meaning, and $w^L(k) = \binom{k}{k_{\mathbf{b}}^L} p(k) \frac{\Gamma(k\alpha)}{\Gamma(n^L + k\alpha)}$, and $g^L(k) = \frac{1}{(k\alpha+n^L)(k\alpha+n^L+1)}$.

These formulas can be directly used to obtain the conditional expectations of the last three rows of Table 8.2. In a very similar way one can easily obtain the conditional expectations contained in the first three rows.

Part III

Discussion

Chapter 9

Discussion and conclusion

This thesis is the result of a five-year study conducted in between the Faculty of Criminal Justice of Lausanne and the Mathematical Institute of Leiden. This hybrid background gave rise to a research that contributed to improving both the domains.

9.1 Contribution to the practice of Forensic Science

The original aim of the thesis was to develop a Bayesian evaluative framework for the results obtained with DIP-STR technology, which in turns constitutes an answer to the problem of extremely unbalanced mixtures. Based on the use of graphical models, it allows using the results obtained with the DIP-STR technology in a legal context, inasmuch it leads to the calculation of the likelihood ratio for any observation.

Additionally, the thesis provides several solutions to deal with the rare type match problem. Generalizations of the Good-Turing estimator and of the discrete Laplace models are used in a frequentist context. A method based on the use of a Bayesian nonparametric prior, and a revisiting of the classical Dirichlet-multinomial model, are proposed and discussed in a Bayesian framework. At the best of our knowledge, this thesis constitutes the first introduction of Bayesian nonparametric prior to forensic applications. Given the satisfying results, we are confident this is only the beginning.

The lemma described in Section 2.5 and proved in Chapter 7 is of very broad application and very useful in many forensic cases. In fact, it allows one to simplify the calculation of the likelihood ratio in all the situations in which prosecution and defence agree on the distribution of part of the available data (for instance, they both see the reference database as a random sample from the population), but they disagree on the distribution of the rest of the data (for instance, the correspondence between the DNA profile of the suspect and of the crime stain is a random event according to the defence, while it is a sure event according to the prosecution).

9.2 Contribution to the Philosophical point of view

Bayesianism dates back of at least one century with, among others, De Finetti (1931, 1937), Wald (1949), and Savage (1954). It is commonly perceived as an opponent to classical frequentist statistics. The use of Bayesian statistics is, according to many, the natural choice for a statistician working in the legal framework, as testified by important pieces of literature such as Lindley (1977b,a, 1978), Robertson and Vignaux (1995), and Aitken and Taroni (2004).

The use of frequentist methods to assess the likelihood ratio may be seen as less coherent, since the likelihood ratio is then used within the Bayes' theorem context, as the way to update prior odds to posterior odds. However, many frequentists statisticians are interested as well in the likelihood ratio, seen as a tool to measure the evidential value of data, independently of the Bayes' theorem.

In order to study the rare type match problem, we came in contact with both the Bayesian and the frequentist approaches to likelihood ratio assessment. We believe that often literature proposes hybrid solutions, passed off as Bayesian, such as the plug-in Bayesian approximations, which in fact can be seen as a compound of the two approaches. Hence, we felt the need to set up a formal distinction between the two. In particular, we wanted to emphasise that the frequentist approach can be seen as a Bayesian approach with special prior over the nuisance parameters. The difference among the two lies in the definition of the probability.

Lastly, the Bayesian plug-in method proposed in much forensic literature, which consists of estimating the unknown allelic frequencies (nuisance parameter of the model) using the mean of their posterior distribution after the observation of a database, is discussed and compared with the full Bayesian approach that integrates out the nuisance parameters. The latter is often not more difficult, but most of the time the likelihood ratios using the two methods do not differ substantially. The Lemma proposed allows one to calculate the full Bayesian likelihood ratios by calculating the posterior expectation of a simple function of the parameter, instead of performing classical marginalization steps.

9.3 Future perspective

The interpretative framework developed for DIP-STR results from a mixed trace, took as assumption that the number of contributors is known, and equal to two. However, DIP-STR markers can also be used to investigate situations in which a fixed number of contributors cannot be agreed with certainty, extending the modular assessment procedures. The uncertainty about the number of contributors is to be taken into account as a further variable when evaluating DIP-STR profiling results. The question of how many individuals have contributed to a given mixture is a general issue that is independent of the type of analysis which is chosen (i.e., traditional STR or DIP-STR), but the fact that DIP-STR alleles have a two-dimensional set of labels for the alleles (two different possible DIP alleles for each STR alleles), could potentially provide more discriminative power over the number of contributors. This can be explored extending the model derived in the first part of the research (the

OOBN constructed to model DIP-STR results), taking stands on the structure proposed in Biedermann et al. (2011b). Through simulations, it is conceivable to compare this method against the classical methods (typically STR and Y-STR) to see if there is an advantage in using DIP-STR markers when the number of contributors in not known in advance.

Despite the numerous results obtained we felt that much more can be done for the rare type match problem. Especially regarding Bayesian nonparametric methodologies, we are just at the dawn: new kinds of nonparametric priors can be studied and used instead of the two-parameter Poisson-Dirichlet distribution, new methods to choose their hyperparameters can be adopted. Of particular interest for the rare type match problem, is the 'k-method' of Brenner (2010). We believe that a rigorous definition of the statistical background is missing, and one of the next steps for this research could be that of better formalising this solution.

9.4 Conclusion

This thesis represented an opportunity to study many crucial problems concerning forensic statistics. This five-year study uncovered many interconnections between them, but there is no clear-cut answer, applicable to all problems. Each one needs a tailored solution, and many aspects should be taken into account and weighed. Issues such as data reduction, uncertainty assessment, hybrid approximations, are rarely discussed and studied in literature. Moreover, there is a big contraposition of Bayesian and frequentist approaches to probability definition. Both approaches have advantages and disadvantages and both need further refinement and improvement. However, the Bayesian approach is more appropriated for legal and forensic reasoning (and easier to explain to lay-persons), even though the choice of the priors is a delicate and crucial problem, often underestimated.

Bibliography

Aitken, C. and Gammerman, A. (1989). Probabilistic reasoning in evidential assessment. *Journal of the Forensic Science Society*, 29:303–316.

Aitken, C. and Taroni, F. (2004). *Statistics and the Evaluation of Evidence for Forensics Scientists*. John Wiley & Sons, Chichester.

Aitken, C., Taroni, F., Barnett, P. D., and Tsatsakis, A. M. (1998). A verbal scale for the interpretation of evidence. *Science & Justice*, 38:279–283.

Aldous, D. J. (1985). *Exchangeability and Related Topics*, volume 1117 of *École D'Été de Probabilités de Saint-Flour*. Springer-Verlag, New York.

Andersen, M. M. (2013). A gentle introduction to the discrete Laplace method for estimating Y-STR haplotype frequencies. arXiv:1304.2129.

Andersen, M. M., Caliebe, A., Jochens, A., Willuweit, S., and Krawczak, M. (2013a). Estimating trace-suspect match probabilities for singleton Y-STR haplotypes using coalescent theory. *Forensic Science International: Genetics*, 7:264–271.

Andersen, M. M., Eriksen, P. S., and Morling, N. (2013b). The discrete Laplace exponential family and estimation of Y-STR haplotype frequencies. *Journal of Theoretical Biology*, 329:39–51.

Anevski, D., Gill, R. D., and Zohren, S. (2013). Estimating a probability mass function with unknown labels. arXiv:1312.1200.

Applied Biosystems (2012). *AmpFlSTR Profiler Plus PCR Amplification Kit User's Manual*. Foster City, California.

Balding, D. (2005). *Weight-of-evidence for Forensic DNA Profiles*. John Wiley & Sons, Chichester.

Balding, D. J. (1995). Estimating products in forensic identification using DNA profiles. *Journal of the American Statistical Association*, 90:839–844.

Ballantyne, K. N., Keerl, V., Wollstein, A., Choi, Y., Zuniga, S. B., Ralf, A., Vermeulen, M., de Knijff, P., and Kayser, M. (2012). A new future of forensic Y-chromosome analysis: rapidly mutating Y-STRs for differentiating male relatives and paternal lineages. *Forensic Science International: Genetics*, 6:208–218.

Bangsø, O. and Wuillemin, P. H. (2000). Object Oriented Bayesian Networks A Framework for Topdown Specification of Large Bayesian Networks and Repetitive Structures. Technical report, Hewlett-Packard Laboratory for Normative Systems, Aalborg University.

Barger, K. and Bunge, J. (2010). Objective Bayesian estimation for the number of species. *Bayesian Analysis*, 5:765–785.

Berger, C. E. and Slooten, K. (2016). The LR does not exist. *Science & Justice*, 56:388–391.

Biedermann, A. (2007). *Bayesian Networks and the Evaluation of Scientific Evidence in Forensic Science*. PhD thesis, University of Lausanne.

Biedermann, A., Garbolino, P., and Taroni, F. (2013). The subjectivist interpretation of probability and the problem of individualisation in forensic science. *Science & Justice*, 53:192–200.

Biedermann, A. and Taroni, F. (2012). Bayesian networks for evaluating forensic DNA profiling evidence: A review and guide to literature. *Forensic Science International: Genetics*, 6:147–157.

Biedermann, A., Taroni, F., Bozza, S., and Aitken, C. (2008). Analysis of sampling issues using Bayesian networks. *Law, Probability and Risk*, 7:35–60.

Biedermann, A., Taroni, F., Bozza, S., and Mazzella, W. D. (2011a). Implementing statistical learning methods through Bayesian networks (part 2): Bayesian evaluations for results of black toner analyses in forensic document examination. *Forensic Science International*, 204:58–66.

Biedermann, A., Taroni, F., and Thompson, W. C. (2011b). Using graphical probability analysis (Bayes Nets) to evaluate a conditional DNA inclusion. *Law, Probability and Risk*, 10:89–121.

Brandwein, A. C. and Strawderman, W. (2005). Bayesian estimation of multivariate location parameters. In *Handbook of Statistics*, volume 25, pages 221–244. Elsevier, Amsterdam.

Brenner, C. H. (2010). Fundamental problem of forensic mathematics – The evidential value of a rare haplotype. *Forensic Science International: Genetics*, 4:281–291.

Brenner, C. H. (2014). Understanding Y haplotype matching probability. *Forensic Science International: Genetics*, 8:233–243.

Brümmer, N. and Swart, A. (2014). Bayesian calibration for forensic evidence reporting. arXiv:1403.5997.

Buckleton, J. and Curran, J. (2005). Sampling effects. In Buckleton, J., Triggs, C., and Walsh, S. J., editors, *Forensic DNA evidence interpretation*, chapter 6, pages 197–216. CRC Press, Boca Raton.

Buckleton, J., Krawczak, M., and Weir, B. (2011). The interpretation of lineage markers in forensic DNA testing. *Forensic Science International: Genetics*, 5:78–83.

Buckleton, J., Triggs, C., and Walsh, S. (2005). *Forensic DNA Evidence Interpretation*. CRC Press, Boca Raton.

Budowle, B., Ge, J., and Chakraborty, R. (2007). Basic principles for estimating the rarity of Y-STR haplotypes derived from forensic evidence. In *18th International Symposium on Human Identification*, Hollywood, CA.

Buntine, W. and Hutter, M. (2010). A Bayesian view of the Poisson-Dirichlet process. arXiv:1007.0296.

Butler, J. (2005). *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers*. Elsevier Academic Press, New York.

Butler, J. (2011). *Advanced Topics in Forensic DNA Typing: Methodology*. Elsevier Academic Press, New York.

Caliebe, A., Jochens, A., Krawczak, M., and Rösler, U. (2010). A markov chain description of the stepwise mutation model: local and global behaviour of the allele process. *Journal of Theoretical Biology*, 266:336–342.

Carlton, M. A. (1999). *Applications of the Two-Parameter Poisson-Dirichlet Distribution*. PhD thesis, University of California, Los Angeles.

Carracedo, A., Bar, W., Lincoln, P., Mayr, W., Morling, N., Olaisen, B., Schneider, P., Budowle, B., Brinkmann, B., Gill, P., Holland, M., Tully, G., and Wilson, M. (2000). DNA commission of the international society for forensic genetics: guidelines for mitochondrial DNA typing. *Forensic Science International*, 110(2):79–85.

Castella, V., Gervaix, J., and Hall, D. (2013). DIP-STR: Highly sensitive markers for the analysis of unbalanced genomic mixtures. *Human Mutation*, 34:644–654.

Cavallini, D. and Corradi, F. (2006). Forensic identification of relatives of individuals included in a database of DNA profiles. *Biometrika*, 93:525–536.

Cereda, G. (2016a). Bayesian approach to LR for the rare match problem. *Statistica Neerlandica*, In Press.

Cereda, G. (2016b). Impact of model choice on LR assessment in case of rare haplotype match (frequentist approach). *Scandinavian Journal of Statistics*, In Press.

Cereda, G. (2016c). Nonparametric Bayesian approach to LR assessment in case of rare haplotype match. arXiv:1506.08444.

Cereda, G., Biedermann, A., Hall, D., and Taroni, F. (2014a). An investigation of the potential of DIP-STR markers for DNA mixture analyses. *Forensic Science International: Genetics*, 11:229–240.

Cereda, G., Biedermann, A., Hall, D., and Taroni, F. (2014b). Object-oriented Bayesian networks for evaluating DIP-STR profiling results from unbalanced DNA mixtures. *Forensic Science International: Genetics*, 8:159–169.

Cereda, G., Gill, R. D., and Taroni, F. (2016). A solution for the rare type match problem when using DIP-STR marker system. Submitted to Forensic Science International: Genetics.

Cerquetti, A. (2010). Bayesian nonparametric analysis for a species sampling model with finitely many types. arXiv:1001.0245.

Chakraborty, R., Stivers, D., Su, B., Zhong, Y., and Budowle, B. (1999). The utility of short tandem repeat loci beyond human identification: Implications for development of new DNA typing systems. *Electrophoresis*, 20:1682–1696.

Champod, C., Biedermann, A., Vuille, J., S., W., and J., D. K. (2016). ENFSI guideline for evaluative reporting in forensic science: A primer for legal practitioners. *Criminal Law & Justice Weekly*, 180(189-193).

Chao, A. and Lee, S.-M. (1992). Estimating the number of classes via sample coverage. *Journal of the American Statistical Association*, 87:210–217.

Chen, Z. and McGee, M. (2008). A Bayesian approach to zero numerator problems using hierarchical models. *Journal of Data Science*, 6:261–268.

Clayton, T. and Buckleton, J. (2005). Mixtures. In Buckleton, J., Triggs, C., and Walsh, S. J., editors, *Forensic DNA Evidence Interpretation*, chapter 7, pages 217–274. CRC Press, Boca Raton.

Cook, R., Evett, I., Jackson, G., Jones, P., and Lambert, J. (1998). A hierarchy of propositions: deciding which level to address in casework. *Science & Justice*, 38:231–239.

Cooper, D. and Krawczak, M. (1991). Mechanisms of insertional mutagenesis in human genes causing genetic disease. *Human Genetics*, 87:409–415.

Coquoz, R. and Taroni, F. (2006). *Preuve par l'ADN: La Génétique au Gervice de la Justice*. Sciences forensiques. Presses polytechniques et universitaires romandes, Lausanne.

Cowell, R., Dawid, P., Lauritzen, S., and Spiegelhalter, D. (2007a). *Probabilistic Networks and Expert Systems: Exact Computational Methods for Bayesian Networks*. Springer, New York.

Cowell, R., Lauritzen, S., and Mortera, J. (2006a). MAIES: A Tool for DNA Mixture Analysis. In *22nd Conference on Uncertainty in Artificial Intelligence*, pages 90–97, San Francisco.

Cowell, R., Lauritzen, S., and Mortera, J. (2006b). Object-oriented Bayesian networks for DNA mixture analyses. (Manuscript available at http://www.staff.city.ac.uk/rgc).

Cowell, R., Lauritzen, S., and Mortera, J. (2011). Probabilistic expert systems for handling artifacts in complex DNA mixtures. *Forensic Science International: Genetics*, 5:202–209.

Cowell, R. G. (2009). Validation of an STR peak area model. *Forensic Science International: Genetics*, 3:193–199.

Cowell, R. G., Lauritzen, S. L., and Mortera, J. (2007b). Identification and separation of DNA mixtures using peak area information. *Forensic Science International*, 166:28–34.

Cowell, R. G., Lauritzen, S. L., and Mortera, J. (2008). Probabilistic modelling for DNA mixture analysis. *Forensic Science International: Genetics Supplement Series*, 1:640–642.

Curran, J., Buckleton, J., Triggs, C., and Weir, B. (2002). Assessing uncertainty in DNA evidence caused by sampling effects. *Science & Justice*, 42:29–37.

Curran, J. M. (2005). An introduction to Bayesian credible intervals for sampling error in DNA profiles. *Law, Probability and Risk*, 4:115–126.

Curran, J. M. (2016). Admitting to uncertainty in the LR. *Science & Justice*, 56:380–382.

da Costa Francez, P. A., Ribeiro Rodrigues, E. M., de Velasco, A. M., and Batista dos Santos, S. E. (2012). Insertion-deletion polymorphisms–utilization on forensic analysis. *International Journal of Legal Medicine*, 126:491–496.

Dawid, A. P. (2001). Comment on Stockmarr's "Likelihood ratios for evaluating DNA evidence when the suspect is found through a database search". *Biometrics*, 57:976–980.

Dawid, A. P. (2016). Forensic lkelihood ratio: statistical problems and pitfalls.

Dawid, A. P. and Evett, I. (1997). Using a graphical method to assist the evaluation of complicated patterns of evidence. *Journal of Forensic Sciences*, 42:226–231.

Dawid, A. P. and Mortera, J. (1996). Coherent analysis of forensic identification evidence. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58:425–443.

Dawid, A. P., Mortera, J., Pascali, V., and Van Boxel, D. (2002). Probabilistic expert systems for forensic inference from genetic markers. *Scandinavian Journal of Statistics*, 29:577–595.

Dawid, A. P., Mortera, J., and Vicard, P. (2007). Object-oriented Bayesian networks for complex forensic DNA profiling problems. *Forensic Science International*, 169:195–205.

Dawid, A. P., Van Boxel, D., Mortera, J., and Pascali, V. L. (1999). Inference about disputed paternity from an incomplete pedigree using a probabilistic expert system. *Bulletin of the International Statistical Institute*, 58:241–242.

De Finetti, B. (1931). Sul significato soggettivo della probabilità. *Fundamenta Mathematicae*, 17:298–329.

De Finetti, B. (1937). La prévision, ses lois logiques, ses sources subjectives. *Ann. Inst. Henri Poincaré*, 7:1–68.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39:1–38.

Egeland, T. and Salas, A. (2008). Estimating haplotype frequency and coverage of databases. *PLoS ONE*, 3:e3988–e3988.

Engen, S. (1975). A note on the geometric series as a species frequency model. *Biometrika*, 62:694–699.

Essen-Möller, E. (1938). Die Beweiskraft der Ähnlichkeit im Vaterschaftsnachweis – Theoretische Grundlagen. *Mitteilungen der Anthropologischen Gesellschaft*, 68:9–53.

Evett, I., Gill, P., Jackson, G., Whitaker, J., and Champod, C. (2002). Interpreting small quantities of DNA: The hierarchy of propositions and the use of Bayesian networks. *Journal of Forensic Sciences*, 47:520–530.

Evett, I., Jackson, G., Lambert, J., and McCrossan, S. (2000). The impact of the principles of evidence interpretation on the structure and content of statements. *Science & Justice*, 40:233–239.

Evett, I. and Weir, B. (1998). *Interpreting DNA evidence: Statistical Genetics for Forensic Scientists*. Sinauer Associates, Sunderland.

Evett, I. W. and Buckleton, J. S. (1996). Statistical analysis of STR data. In *Advances in Forensic Haemogenetics*. Springer Verlag.

Evett, I. W., Buffery, C., Willott, G., and Stoney, D. (1991). A guide to interpreting single locus profiles of DNA mixtures in forensic cases. *Journal of the Forensic Science Society*, 31:41–47.

Ewens, W. (1972). Sampling theory of selectively neutral alleles. *Theoretical Population Biology*, 3:87–112.

Favaro, S., Lijoi, A., Mena, R. H., and Pruenster, I. (2009). Bayesian nonparametric inference for species variety with a two parameter Poisson-Dirichlet process prior. *Journal of the Royal Statistical Society: Series B (Methodological)*, 71:993–1008.

Feng, S. (2010). *The Poisson-Dirichlet Distribution and Related Topics: Models and Asymptotic Behaviors*. Springer, Berlin.

Fenton, N. and Neil, M. (2012). *Risk Assessment and Decision Analysis with Bayesian Networks*. CRC Press, Boca Raton.

Foreman, L., Smith, A., and Evett, I. (1997). Bayesian analysis of DNA profiling data in forensic identification applications. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 160:429–459.

Gadi, V., Nelson, J., Boespflug, N., Guthrie, K., and Kuhr, C. (2006). Soluble donor DNA concentrations in recipient serum correlate with pancreas-kidney rejection. *Clinical Chemistry*, 52:379–382.

Gale, W. A. and Church, K. W. (1994). What's wrong with adding one? In *Corpus-Based Research into Language. Rodolpi*.

Garbolino, P. and Taroni, F. (2002). Evaluation of scientific evidence using Bayesian networks. *Forensic Science International*, 125:149–155.

Gill, P., Brenner, C., Brinkmann, B., Budowle, B., Carracedo, A., Jobling, M. A., de Knijff, P., Kayser, M., Krawczak, M., Mayr, W. R., Morling, N., Olaisen, B., Pascali, V., Prinz, M., Roewer, L., Schneider, P. M., Sajantila, A., and Tyler-Smith, C. (2001). DNA Commission of the International Society of Forensic Genetics: recommendations on forensic analysis using Y-chromosome STRs. *Forensic Science International*, 124:5–10.

Gittelson, S., Biedermann, A., Bozza, S., and Taroni, F. (2012). The database search problem: A question of rational decision making. *Forensic Science International*, 222:186–199.

Gnedin, A. (2009). A Species Sampling Model with Finitely many Types. arXiv:0910.1988.

Gnedin, A., Hansen, B., and Pitman, J. (2007). Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. *Probability Surveys*, pages 146–171.

Gnedin, A. and Pitman, J. (2006). Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical Sciences*, 138:5674–5685.

Gómez, M. (2004). Real-World Applications of Influence Diagrams. In Gámez, J. A., Moral, S., and Salmerón, A., editors, *Advances in Bayesian Networks*. Springer, Berlin.

Good, I. (1950). *Probability and the Weighing of Evidence*. Charles Griffin, London.

Good, I. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40:237–264.

Gorenflo, R., Kilbas, A., Mainardi, F., and Rogosin, S. (2014). *Mittag-Leffler Functions, Related Topics and Applications*. Springer Monographs in Mathematics. Springer Berlin Heidelberg.

Green, P. J. and Mortera, J. (2009). Sensitivity of inferences in forensic genetics to assumptions about founding genes. *Annals of Applied Statistics*, 3:731–763.

Green, R. L., Lagace, R. E., Oldroyd, N. J., Hennessy, L. K., and Mulero, J. J. (2013). Developmental validation of the AmpFℓSTR®NGM SElect^{TM}PCR Amplification Kit: A next-generation STR multiplex with the SE33 locus. *Forensic Science International: Genetics*, 7:41–51.

Gunel, E. and Wearden, S. (1995). Bayesian estimation and testing of gene frequencies. *Theoretical and Applied Genetics*, 91:534–543.

Guo, X., Bayliss, P., Damewood, M., Varney, J., Ma, E., Vallecillo, B., and Dhallan, R. (2012). A noninvasive test to determine paternity in pregnancy. *New England Journal of Medicine*, 366:1743–1745.

Haas, C., Wangensteen, T., Giezendanner, N., Kratzer, A., and Bär, W. (2006). Y-chromosome STR haplotypes in a population sample from Switzerland (Zurich area). *Forensic Science International*, 158:213–218.

Haas, P. J. and Stokes, L. (1998). Estimating the number of classes in a finite population. *Journal of the American Statistical Association*, 93:1475–1487.

Hepler, A. B., Dawid, A. P., and Leucari, V. (2007). Object-oriented graphical representations of complex patterns of evidence. *Law, Probability and Risk*, 6:275–293.

Hepler, A. B. and Weir, B. S. (2008). Object-oriented Bayesian networks for paternity cases with allelic dependencies. *Forensic Science International: Genetics*, 2:166–175.

Hill, B. M. (1968). Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *Journal of the American Statistical Association*, 63:677–691.

Hill, B. M. (1979). Posterior moments of the number of species in a finite population and the posterior probability of finding a new species. *Journal of the American Statistical Association*, 74:668–673.

Hjort, N., Holmes, C., Müller, P., and Walker, S. (2010). *Bayesian Nonparametrics*. Cambridge University Press, Cambridge.

Inusah, S. and Kozubowski, T. (2006). A discrete analogue of the Laplace distribution. *Journal of Statistical Planning and Inference*, 136:1090–1102.

Jensen, F. and Nielsen, T. (2007). *Bayesian Networks and Decision Graphs*. Springer, New York.

Jobling, M. A. and Gill, P. (2004). Encoded evidence: DNA in forensic analysis. *Nature Reviews Genetics*, 5:739–751.

Jordan, M. (1998). *Learning in Graphical Models*. MIT Press, Cambridge.

Jordan, M. I. (2004). Graphical models. *Statistical Science*, 19:140–155.

Jovanovic, B. D. and Levy, P. S. (1997). A look at the rule of three. *The American Statistician*, 51:137–139.

Karlin, S. (1967). Central limit theorems for certain infinite urn schemes. *Journal of Mathematics and Mechanics*, 17:373–401.

Karlin, S. and McGregor, J. (1972). Addendum to a paper of W. Ewens. *Theoretical Population Biology*, 3:113–116.

Kaufman, L. and Rousseeuw, P. (2009). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, New Jersey.

Kimura, M. (1964). The number of alleles that can be maintained in a finite population. *Genetics*, 49:725–738.

Kimura, M. and Ohta, T. (1978). Stepwise mutation model and distribution of allelic frequencies in a finite population. *Proceedings of the National Academy of Sciences of the United States of America*, 75:2868–2872.

Kingman, J. (1977). The population structure associated with the Ewens sampling formula. *Theoretical Population Biology*, 11:274–283.

Kingman, J. (1978). Random partitions in population-genetics. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 361:1–20.

Kingman, J. (1980). *Mathematics of Genetic Diversity*. Society for Industrial and Applied Mathematics, Philadelphia.

Kingman, J. F. C. (1975). Random discrete distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 37:1–22.

Kjærulff, U. B. and Madsen, A. L. (2008). *Bayesian Networks and Influence Diagrams, A Guide to Construction and Analysis*. Springer, New York.

Koller, D. and Pfeffer, A. (1997). Object-oriented Bayesian networks. In *Proceedings of the 13th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 302–313.

Konis, K. (2010). *RHugin: RHugin*. R package version 7.4/r247.

Korb, K. and Nicholson, A. (2011). *Bayesian Artificial Intelligence*. CRC Press, Boca Raton.

Koski, T. and Noble, J. (2011). *Bayesian Networks: An Introduction*. John Wiley & Sons, Chichester.

Krawczak, M. (2001). Forensic evaluation of Y-STR haplotype matches: a comment. *Forensic Science International*, 118:114–115.

Krichevsky, R. and Trofimov, V. (1981). The performance of universal coding. *IEEE Transactions on Information Theory*, 27:199–207.

Lange, K. (1995). Applications of the Dirichlet distribution to forensic match probabilities. *Genetica*, 96:107–117.

Laplace, P. (1814). *Essai Philosophique sur les Probabilites*. Mme. Ve Courcier, Paris.

Laskey, K. and Mahoney, S. (1997). Network fragments: Representing knowledge for constructing probabilistic models. In *UAI'97 Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*, pages 334–341.

Lewins, W. A. and Joanes, D. N. (1984). Bayesian estimation of the number of species. *Biometrics*, 40:323–328.

Lijoi, A., Mena, R. H., and Pruenster, I. (2007). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika*, 94:769–786.

Lindley, D. (1977a). Probability and the law. *The Statistician*, 26:203–220.

Lindley, D. (1977b). A problem in forensic science. *Biometrika*, 64:207–213.

Lindley, D. (1978). The Bayesian approach. *Scandinavian Journal of Statistics*, 5:1–26.

Lindley, D. (1991). Subjective probability, decision analysis and their legal consequences. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 154:83–92.

Lo, Y., Tein, M., Lau, T., Haines, C., Leung, T., Poon, P., Wainscoat, J., Johnson, P., Chang, A., and Hjelm, N. (1998). Quantitative analysis of fetal DNA in maternal plasma and serum: Implications for noninvasive prenatal diagnosis. *American Journal of Human Genetics*, 62:768–775.

Louis, T. A. (1981). Confidence intervals for a binomial parameter after observing no successes. *The American Statistician*, 35:154–154.

Lucy, D. (2005). *Introduction to Statistics for Forensic Scientists*. John Wiley & Sons, Chichester.

Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S., and Devine, S. E. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Research*, 16:1182–1190.

Morrison, G. S. (2010). *Evidence Expert*, chapter Forensic voice comparison. Thomson Reuters, Sidney, Australia.

Mortera, J., Dawid, A. P., and Lauritzen, S. L. (2003). Probabilistic expert systems for DNA mixture profiling. *Theoretical Population Biology*, 63:191–205.

Neapolitan, R. E. (1990). *Probabilistic Reasoning in Expert Systems*. John Wiley & Sons, Inc., New York.

Neuvonen, A. M., Palo, J. U., Hedman, M., and Sajantila, A. (2012). Discrimination power of investigator DIPplex loci in Finnish and Somali populations. *Forensic Science International: Genetics*, 6:e99 – e102.

Newman, M. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46:323–351.

Oldoni, F., Castella, V., and Hall, D. (2015). A novel set of DIP-STR markers for improved analysis of challenging DNA mixtures. *Forensic Science International: Genetics*, 18:156–164.

Orlitsky, A., Santhanam, N., and Zhang, J. (2003). Always Good Turing: asymptotically optimal probability estimation. *Science*, 302:427–431.

Orlitsky, A., Santhanam, N. P., Viswanathan, K., and Zhang, J. (2004). On Modeling Profiles Instead of Values. In *Uncertainty in Artificial Intelligence*, pages 426–435.

Pearl, J. (1982). Reverend Bayes on inference engines: A distributed hierarchical approach. In *Proceedings of the National Conference on Artificial Intelligence (AAAI'82)*, pages 133–136, Pittsburgh. Morgan Kaufmann.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. The Morgan Kaufmann Series in Representation and Learning. Morgan Kaufmann Publisher, Inc., San Mateo.

Pereira, R., Phillips, C., Alves, C., Amorim, A., Carracedo, A., and Gusmao, L. (2009a). A new multiplex for human identification using insertion deletion polymorphisms. *Electrophoresis*, 30:3682–3690.

Pereira, R., Phillips, C., Alves, C., Amorim, A., Carracedo, Á., and Gusmão, L. (2009b). Insertion/deletion polymorphisms: A multiplex assay and forensic applications. *Forensic Science International: Genetics Supplement Series*, 2:513 – 515.

Perman, M., Pitman, J., and Yor, M. (1992). Size-biased sampling of Poisson point-processes and excursions. *Probability Theory and Related Fields*, 92:21–39.

Pitman, J. (1992). The two-parameter generalization of Ewens' random partition structure. Technical report 345, Department of Statistics U.C. Berkeley CA.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102:145–158.

Pitman, J. (1996). Some Developments of the Blackwell-MacQueen Urn Scheme. In *Lecture Notes-Monograph Series*, volume 30, pages 245–267.

Pitman, J. and Picard, J. (2006). *Combinatorial Stochastic Processes*. École D'Été de Probabilités de Saint-Flour XXXII - 2002. Springer, Berlin.

Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25:855–900.

Pourret, O., Naïm, P., and Marcot, B. (2008). *Bayesian Networks: A Practical Guide to Applications*. Statistics in Practice. John Wiley & Sons, Chichester.

Press, S. (2009). *Subjective and Objective Bayesian Statistics: Principles, Models, and Applications*. Wiley Series in Probability and Statistics. Wiley, Hoboken, New Jersey.

Pujal, J.-M. and Gallardo, D. (2008). PCR-based methodology for molecular microchimerism detection and quantification. *Experimental Biology and Medicine*, 233:1161–1170.

Purps, J., Siegert, S., Willuweit, S., Nagy, M., Alves, C., Salazar, R., Angustia, S. M. T., Santos, L. H., Anslinger, K., Bayer, B., Ayub, Q., Wei, W., Xue, Y., Tyler-Smith, C., Bafalluy, M. B., Martínez-Jarreta, B., Egyed, B., Balitzki, B., Tschumi, S., Ballard, D., Court, D. S., Barrantes, X., Bäßler, G., Wiest, T., Berger, B., Niederstätter, H., Parson, W., Davis, C., Budowle, B., Burri, H., Borer, U., Koller, C., Carvalho, E. F., Domingues, P. M., Chamoun, W. T., Coble, M. D., Hill, C. R., Corach, D., Caputo, M., D'Amato, M. E., Davison, S., Decorte, R., Larmuseau, M. H. D., Ottoni, C., Rickards, O., Lu, D., Jiang, C., Dobosz, T., Jonkisz, A., Frank, W. E., Furac, I., Gehrig, C., Castella, V., Grskovic, B., Haas, C., Wobst, J., Hadzic, G., Drobnic, K., Honda, K., Hou, Y., Zhou, D., Li, Y., Hu, S., Chen, S., Immel, U.-D., Lessig, R., Jakovski, Z., Ilievska, T., Klann, A. E., García, C. C., de Knijff, P., Kraaijenbrink, T., Kondili, A., Miniati, P., Vouropoulou, M., Kovacevic, L., Marjanovic, D., Lindner, I., Mansour, I., Al-Azem, M., Andari, A. E., Marino, M., Furfuro, S., Locarno, L., Martín, P., Luque, G. M., Alonso, A., Miranda, L. S., Moreira, H., Mizuno, N., Iwashima, Y., Neto, R. S. M., Nogueira, T. L. S., Silva, R., Nastainczyk-Wulf, M., Edelmann, J., Kohl, M., Nie, S., Wang, X., Cheng, B., Núñez, C., Pancorbo, M.

M. d., Olofsson, J. K., Morling, N., Onofri, V., Tagliabracci, A., Pamjav, H., Volgyi, A., Barany, G., Pawlowski, R., Maciejewska, A., Pelotti, S., Pepinski, W., Abreu-Glowacka, M., Phillips, C., Cárdenas, J., Rey-Gonzalez, D., Salas, A., Brisighelli, F., Capelli, C., Toscanini, U., Piccinini, A., Piglionica, M., Baldassarra, S. L., Ploski, R., Konarzewska, M., Jastrzebska, E., Robino, C., Sajantila, A., Palo, J. U., Guevara, E., Salvador, J., Ungria, M. C. D., Rodriguez, J. J. R., Schmidt, U., Schlauderer, N., Saukko, P., Schneider, P. M., Sirker, M., Shin, K.-J., Oh, Y. N., Skitsa, I., Ampati, A., Smith, T.-G., Calvit, L. S. d., Stenzl, V., Capal, T., Tillmar, A., Nilsson, H., Turrina, S., De Leo, D., Verzeletti, A., Cortellini, V., Wetton, J. H., Gwynne, G. M., Jobling, M. A., Whittle, M. R., Sumita, D. R., Wolańska-Nowak, P., Yong, R. Y. Y., Krawczak, M., Nothnagel, M., and Roewer, L. (2014). A global analysis of Y-chromosomal haplotype diversity for 23 STR loci. *Forensic Science International: Genetics*, 12:12–23.

Reynolds, R., Sensabaugh, G., and Blake, E. (1991). Analysis of genetic-markers in forensic DNA samples using the polymerase chain-reaction. *Analytical Chemistry*, 63:2–15.

Robertson, B. and Vignaux, G. A. (1995). *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*. John Wiley & Sons, Chichester.

Roewer, L. (2009). Y chromosome STR typing in crime casework. *Forensic Science, Medicine, and Pathology*, 5:77–84.

Roewer, L., Amemann, J., Spurr, N. K., Grzeschik, K. H., and Epplen, J. T. (1992). Simple repeat sequences on the human Y-chromosome are equally polymorphic as their autosomal counterparts. *Human Genetics*, 89:389–394.

Roewer, L., Kayser, M., de Knijff, P., Anslinger, K., Betz, A., Caglia, A., Corach, D., Furedi, S., Henke, L., Hidding, M., Kargel, H., Lessig, R., Nagy, M., Pascali, V., Parson, W., Rolf, B., Schmitt, C., Szibor, R., Teifel-Greding, J., and Krawczak, M. (2000). A new method for the evaluation of matches in non-recombining genomes: application to Y-chromosomal short tandem repeat (STR) haplotypes in European males. *Forensic Science International*, 114:31–43.

Rosenberg, N., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J., and Feldman, M. (2005). Clines, clusters, and the effect of study design on the inference of human population structure. *Plos Genetics*, 1:660–671.

Savage, L. J. (1954). *The Foundations of Statistics*. Wiley, New York.

Schum, D. (1994). *The Evidential Foundations of Probabilistic Reasoning*. Northwestern University Press, Evanston.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464.

Sjerps, M. J., Alberink, I., Bolck, A., Stoel, R. D., Vergeer, P., and van Zanten, J. H. (2016). Uncertainty and LR: to integrate or not to integrate, that's the question. *Law, Probability and Risk*, 15:23–29.

Steele, C. D. and Balding, D. J. (2014). Statistical evaluation of forensic DNA profile evidence. *Annual Review of Statistics and Its Application*, 1:361–384.

Stoel, R. D. and Sjerps, M. (2012). Interpretation of forensic evidence. In *Handbook of Risk Theory*, pages 135–158. Springer Netherlands.

Sutherland, C., O'Brien, R., Figarelli, D., Ring, J., and Grates, K. (2009). Evaluation of eight commercially available STR kits - technology evaluation. *NCJRS Abstract Database: National institute*.

Taroni, F., Aitken, C., Garbolino, P., and Biedermann, A. (2006). *Bayesian Networks and Probabilistic Inference in Forensic Science*. John Wiley & Sons, Chichester.

Taroni, F., Biedermann, A., Bozza, S., Garbolino, P., and Aitken, C. (2014). *Bayesian Networks for Probabilistic Inference and Decision Analysis in Forensic Science*. John Wiley & Sons, Chichester, second edition.

Taroni, F., Biedermann, A., Garbolino, P., and Aitken, C. (2004). A general approach to Bayesian networks for the interpretation of evidence. *Forensic Science International*, 139:5–16.

Taroni, F., Biedermann, A., Vuille, J., and Morling, N. (2013). Whose DNA is this? This is not the relevant question (a note for forensic scientists). *Forensic Science International: Genetics*, 7:467–470.

Taroni, F., Bozza, S., Biedermann, A., and Aitken, C. (2016). Dismissal of the illusion of uncertainty in the assessment of a likelihood ratio. *Law, Probability and Risk*, 0:1–16.

Taroni, F., Bozza, S., Biedermann, A., Garbolino, P., and Aitken, C. (2010). *Data Analysis in Forensic Science: A Bayesian Decision Perspective*. John Wiley & Sons, Chichester.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1581.

Thompson, J. M., Ewing, M. M., Frank, W. E., Pogemiller, J. J., Nolde, C. A., Koehler, D. J., Shaffer, A. M., Rabbach, D. R., Fulmer, P. M., Sprecher, C. J., and Storts, D. R. (2013). Developmental validation of the PowerPlex (R) Y23 System: A single multiplex Y-STR analysis system for casework and database samples. *Forensic Science International: Genetics*, 7:240–250.

Tippett, C. F., Emerson, V. J., Fereday, M. J., Lawton, F., Richardson, A., Jones, L. T., and Lampert, M. S. M. (1968). The evidential value of the comparison of paint flakes from sources other than vehicles. *Journal of the Forensic Science Society*, 8:61–65.

Tiwari, R. C. and Tripathi, R. C. (1989). Nonparametric Bayes estimation of the probability of discovering a new species. *Communications in statistics: Theory and methods*, A18:877–895.

Tjoa, M. L., Cindrova-Davies, T., Spasic-Boskovic, O., Bianchi, D. W., and Burton, G. J. (2006). Trophoblastic oxidative stress and the release of cell-free feto-placental DNA. *American Journal of Pathology*, 169:400–404.

Triggs, C. M. and Curran, J. M. (2006). The sensitivity of the Bayesian HPD method to the choice of prior. *Science & Justice*, 46:169–178.

Vali, U., Brandstrom, M., Johansson, M., and Ellegren, H. (2008). Insertion-deletion polymorphisms (indels) as genetic markers in natural populations. *BMC Genetics*, 9:1–8.

Van der Hout, A. and Alberink, I. (2015). Posterior distributions for likelihood ratios in forensic science. http://www.ucl.ac.uk/ ucakadl/LR2015.pdf.

Vermeulen, M., Wollstein, A., van der Gaag, K., Lao, O., Xue, Y., Wang, Q., Roewer, L., Knoblauch, H., Tyler-Smith, C., de Knijff, P., and Kayser, M. (2009). Improving global and regional resolution of male lineage differentiation by simple single-copy Y-chromosomal short tandem repeat polymorphisms. *Forensic Science International: Genetics*, 3:205–213.

Wald, A. (1949). Statistical decision functions. *The Annals of Mathematical Statistics*, 20:165–205.

Walsh, S. J., Ribaux, O., Buckleton, J. S., Ross, A., and Roux, C. (2004). DNA profiling and criminal justice: A contribution to a changing debate. *Australian Journal of Forensic Sciences*, 36:34–43.

Wambaugh, J. (1989). *The Blooding*. Bantam Books, New York.

Weber, J. L., David, D., Heil, J., Fan, Y., Zhao, C., and Marth, G. (2002). Human Diallelic Insertion/Deletion Polymorphisms. *The American Journal of Human Genetics*, 71:854–862.

Weir, B. (1996). *Genetic Data Analysis 2*. Sinauer Associates, Sunderland.

Weir, B., Triggs, C., Buckleton, J., Walsh, K., Stowell, L., and Starling, L. (1997). Interpreting DNA mixtures. *Journal of Forensic Sciences*, 42:213–222.

Willuweit, S., Caliebe, A., Andersen, M. M., and Roewer, L. (2011). Y-STR frequency surveying method: A critical reappraisal. *Forensic Science International: Genetics*, 5:84–90.

Winkler, R. L., Smith, J. E., and Fryback, G. (2002). The role of informative priors in zero-numerator problems: being conservative versus being candid. *The American Statistician*, 64:1–4.

Yang, N., Li, H., Criswell, L., Gregersen, P., Alarcon-Riquelme, M., Kittles, R., Shigeta, R., Silva, G., Patel, P., Belmont, J., and Seldin, M. (2005). Examination of ancestry and ethnic affiliation using highly informative diallelic DNA markers: application to diverse and admixed populations and implications for clinical epidemiology and forensic medicine. *Human Genetics*, 118:382–392.

Samenvatting

In een justitiële zaak kan de *likelihood ratio* als een statistiek gebruikt worden om een bewijsstuk op waarde te schatten. Namelijk, de mate waarin het bewijsstuk de hypotheses van de aanklager ondersteunt ten opzichte van de hypotheses van de verdediging. Een type bewijsstuk dat vaak gevonden wordt op de plaats van het delict is DNA. Dit proefschrift richt zich op de eigenschappen van de *likelihood ratio* wanneer als bewijsstuk DNA gebruikt wordt.

Het statistisch op waarde schatten van DNA-mengels (sporen met DNA van verschillende personen) is ingewikkeld omdat verschillende combinaties van DNA-profielen van bekende en/of onbekende personen compatibel kunnen zijn met het gevonden DNA-mengsel. Bovendien, als het kwantificeerbare aandeel van de DNA van één van de personen minder is dan 10% van de totale hoeveelheid, dan wordt het DNA-profiel van de persoon "gemaskeerd" door het DNA-profiel van de andere personen / of persoon. Het is erg moeilijk om met de klassieke werkwijzen van genotypering en de standaard aan statistische technieken om het genotype van deze gemaskeerde kleine hoeveelheden van DNA te detecteren. Dergelijke uiterst onevenwichtige mengsels van DNA komen echter vaak voor, bijvoorbeeld in het geval van seksueel geweld waar het DNA-mengsel wordt overheerst door dat van het slachtoffer. Oftewel, er is een dringende vraag naar een oplossing om deze kleine hoeveelheden aan DNA betrouwbaar en gemakkelijker te kunnen detecteren. Met de komst van nieuwe technologie zoals het DIP-STR (*Deletion Insertion Polymorphisms - Short Tandem Repeats*) markersysteem wordt er een oplossing geboden voor dit uiterst onevenwichtige DNA mengsel probleem.

Het oorspronkelijke doel van dit proefschrift, vervuld in hoofdstuk 3, was om een Bayesiaans statistisch model te ontwikkelen dat DIP-STR resultaten op waarde zou kunnen schatten in het licht van de belangrijke en concurrerende hypotheses; een essentieel element voor het weergeven van de potentie van deze nieuwe techniek in de toepassing voor beoefenaars. In hoofdstuk 4, hebben we vanuit een statistisch en forensisch oogpunt, en met betrekking tot toepasbaarheid en gebruikersgemak, de DIP-STR markers vergeleken met die van traditionele markersystemen, zoals klassieke STR en Y-STR markers.

Tijdens de voortgang van ons onderzoek, zijn we enkele delicate methodologische vraagstukken tegenkomen voor de forensische statistiek. Een eerste bevinding was dat wat in de literatuur een Bayesiaanse methode genoemd wordt beter gezien kan worden als een ad hoc benadering van de volledige Bayesiaanse oplossing. Vervolgens kwamen we in aanraking met het zeldzaam type match probleem: de situatie waarin er een match wordt gevonden tussen de kenmerken van bepaalde teruggewonnen DNA-materiaal en die van het DNA-controlemateriaal, dit terwijl bij eerder verzamelde monsters deze match nog niet was waargenomen (dat wil zeggen,

de match was niet aanwezig in de op dat moment beschikbare database). Het zeldzaam type match probleem is in het bijzonder aanwezig in situaties waar gebruik wordt gemaakt van Y-STR (of mitochondriaal) DNA profielen, of wanneer er met de genotyperings-technieken wordt gewerkt, zoals DIP-STR markers, waarvoor de beschikbare database nog maar beperkt is in grootte.

In ons onderzoek richten we ons op de Y-STR data voor het bestuderen van zowel de nieuwe en huidige oplossingen voor het zeldzaam type probleem. In Hoofdstuk 6 hebben we de klassieke Bayesiaanse methoden (met beta-binomiale en Dirichlet-multinomiale verdelingen) herzien, en vervolgens vergeleken met een non-parametrische Bayesiaanse benadering die speciaal was ontwikkeld voor de zeldzaam Y-STR match probleem, zie Hoofdstuk 7).

Twee Frequentistische oplossingen voor het zeldzaam type probleem worden geanalyseerd in hoofdstuk ?efp3: de discrete Laplace-methode en een nieuwe oplossing op basis van de Good-Turing schatter. Tijdens het bestuderen van oplossingen vanuit het Frequentistisch perspectief, zijn we erachter gekomen dat verschillende methodes gebaseerd zijn op data reductie, en dat dat zelden aan bod komt in de Forensische literatuur. Bovendien zijn er verschillende niveaus van onzekerheid welke in acht genomen dienen te worden. Door te werken aan beide de Frequentistische en de Bayesiaanse methodes hebben we het verschil tussen de twee benaderingen beter leren te begrijpen, en het verschil tussen de volledige - en de plug-in Bayesiaanse benadering. Om de volledige Bayesiaanse *likelihood ratio* te verkrijgen onder verschillende regulariteits-condities, hebben we een lemma bewezen.

Ter afsluiting van het project is een van de ontwikkelde Bayesiaanse methodes voor het zeldzaam type match probleem ook toegepast op de DIP-STR data in Hoofdstuk 8. Dit model dat geconstrueerd is voor de DIP-STR data, en nog in haar kinderschoenen staat, is verder verbeterd door het uit te breiden op een manier waarbij het wordt toegestaan om de onzekerheid van de parameters op te nemen in het model op een consistente Bayesiaanse manier.

Acknowledgments

This thesis would not exist without the help of my supervisors. I will thank them in the order we met, having been for me equally important and indispensable.

Thank you Richard, for the brilliant ideas and your technical help, but above all, for your incredible human compassion and understanding. Thanks for teaching me (perhaps not always intentionally) the gift of flexibility, and for training me to react positively to all the small obstacles I could meet.

Thank you Franco for your inexhaustible tenacity, and the enthusiasm with which you welcomed each small result of mine. You gave me the energy, the optimism and the self-esteem I needed in the occasional moments of discouragement. Thanks for being understanding and always available for me, and to have encouraged all my choices for the future.

Thanks you Alex for the important contribution to this thesis. Without your impeccable commitment, and your exceptional attention, this work would have been, for sure, less valiant and precise.

I am grateful to the members of the reading committee for their time spent in studying the manuscript in order to come to their judgement.

Thanks to my family, the harbour I reach after each small shipwreck. Since I was young, you helped me to strengthen my weaknesses and to enhance my qualities. Thanks for the possibility you gave me to concentrate completely on my studies, while you took care of all my necessities, and for letting me free to make my own choices.

Thank you Diego, for being at my side for so long. Thanks for having encouraged my independence and having supported me in this difficult and exciting journey, even though this often meant to be physically far away.

Thank you Melania, for having always been a good friend, sharing all important moments of my life.

Thanks to four young women, Elena, Irene, Johanna, Elisabetta, responsible for the increasing in laughter and for many frivolous moments that compensated for any tiredness and frustrations. Thank you Cecilia for the hugs and the smiles, and for bringing me back to your incredible childish word.

Thanks to all my friends, close and far. There are many, and I will try not to forget those which have something to do with this thesis. Thanks Ilaria, Lara, Laura, Liv, Margaux, Marina,

Sonja, Dirk, Luca, Ricardo, Sthéphanie, Vincent, Fengnan, Maarten (who also helped me out with the Samensvatting section).

Ringraziamenti

Questa tesi non sarebbe la stessa senza il prezioso aiuto dei miei due *supervisors*. Essendo stati indispensabili in egual misura, li ringrazio nell'ordine in cui li ho conosciuti, nel lontano 2011.

Grazie a Richard per avermi aiutato tecnicamente con le tue brillanti doti di matematico, ma soprattutto per esserti aperto con me e avermi mostrato il tuo lato umano, e per la tua paziente e profonda comprensione. Grazie per avermi insegnato, forse involontariamente, il dono della flessibilità e per avermi allenato ad affrontare gli imprevisti in maniera positiva.

Grazie a Franco, per la tua inesauribile tenacia e l'entusiasmo con cui hai accompagnato ogni mio progresso. Per essere stato un supervisor incoraggiante e flessibile, pronto a spenderti in prima persona per i tuoi studenti, disponibile fino all'inverosimile. Grazie per avermi fornito l'energia e l'ottimismo necessari nei vari momenti di scoraggiamento e per avermi supportato nelle mie scelte future.

Grazie Alex per l'importante contributo a questa tesi e ai *papers* di cui sei coautore. Senza il tuo impeccabile e eccezionale impegno ed attenzione questo lavoro sarebbe stato certamente meno valido e preciso.

Grazie ai membri della *reading commettee* per il tempo che hanno dedicato a studiare questo manoscritto e per gli utili commenti che mi hanno fornito.

Grazie alla mia famiglia, il porto dove attracco dopo ogni piccolo naufragio. Da quando sono piccola mi avete aiutato a rafforzarmi nelle mie debolezze e a valorizzare le mie qualità. Grazie per la possibilità che mi avete dato di dedicarmi completamente allo studio, mentre voi vi siete presi cura di tutte le mie necessità. E grazie per avermi lasciata libera in ogni mia scelta, sostenendomi indiscriminatamente.

Grazie ai miei nonni, quelli a cui posso raccontare di questo piccolo successo, e quelli che non ci sono più, ma che sono senza dubbio ugualmente orgogliosi di me.

Grazie a Diego, per essere stato al mio fianco per tutto questo tempo. Grazie per aver sempre incoraggiato la mia indipendenza e sostenuto questo percorso difficile ma così emozionante, nonostante questo abbia significato essere spesso (fisicamente) lontani.

Grazie Melania, per essere stata sempre un'amica presente, che ha condiviso tutti i momenti importanti della mia vita, e ha sempre saputo incoraggiarmi e sostenermi.

Grazie a Elena, Irene, Johanna, Elisabetta, quattro giovani donne che sono state uno dei

miei sostegni principali di questi anni. Vi ringrazio per aver contribuito soprattutto a incrementare il numero di risate e di momenti frivoli e leggeri che compensavano alle fatiche, alle frustrazioni, e alle difficoltà di questo percorso.

Grazie ai miei tantissimi amici, vicini e lontani. Cercherò di non dimenticare quelli che in qualche modo hanno avuto a che fare con questa tesi. Grazie Ilaria, Lara, Liv, Margaux, Marina, Sonja, Stéphanie, Dirk, Luca, Ricardo, Vincent, Fengnan, Maarten (che mi ha anche aiutato a tradurre in olandese la sezione "Samensvatting").

Curriculum Vitae

Giulia Cereda was born in Milano on the 07 April 1988. She attended the Università degli Studi di Milano, completing a Master degree in Mathematics in 2011. In November 2011, she started to work as Phd student at the Faculty of Forensic Science (Université de Lausanne) on a SFNS grant, under the supervision of Prof. Franco Taroni. In 2012 she also became a Phd student at the Mathematical Institute (Universiteit Leiden) under the supervision of Prof. Richard Gill, starting a joint Phd research between the two universities. Primarily based in Lausanne, she spent two periods of six months in Leiden, funded by an additional SFNS Mobility grant.